The Dissertation Committee for Xi Chen
certifies that this is the approved version of the following dissertation:

Mathematical modeling of epidemic surveillance

Committee:

Lauren A. Meyers, Supervisor

John Hasenbein

Purnamrita Sarkar

Peter Mueller

Mathematical modeling of epidemic surveillance

by

Xi Chen,

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2019

Dedicated to my family.

Mathematical modeling of epidemic surveillance

Publication No. _____

Xi Chen, Ph.D.
The University of Texas at Austin, 2019

Supervisor: Lauren A. Meyers

My thesis focus on three aspects of epidemic surveillance: Estimation of the probability and corresponding uncertainty analysis for disease to be imported into multiple geographic regions (Chapter 1); Estimation of the transmission of disease after local transmission established (Chapter 2); Prevalence and corresponding confidence interval estimation incorporating individual level test sensitivity and specificity (Chapter 3).

The maximum entropy model, a commonly used species distribution model (SDM) normally combines observations of the species occurrence with environmental information to predict the geographic distributions of animal or plant species. However, it only produces point estimates for the probability of species existence. To understand the uncertainty of the point estimates, we analytically derived the variance of the outputs of the maximum entropy model from the variance of the input in chapter 1. We applied the analytic method to obtain the standard deviation of dengue importation probability

and Aedes aegypti suitability. Dengue occurrence data and Aedes aegypti mosquito abundance data, combined with demographic and environmental data, were applied to obtain point estimates and the corresponding variance. To address the issue of not having the true distributions for comparison, we compared and contrasted the performance of the analytical expression with the bootstrap method and Poisson point process model which proved of equivalence of maximum entropy model with the assumption of independent point locations. Both Dengue importation probability and Aedes aegypti mosquito suitability examples show that the methods generate comparatively the same results and the analytic method we introduced is dramatically faster than the bootstrap method and directly apply to maximum entropy model.

Infectious diseases such as influenza progress quickly potentially reaching large parts of populations. Accurately estimating the parameters of the infectious disease progression model can efficiently help health organization determine the progression and severity of the disease and response properly and quickly. In chapter 2, we studied the application of 2 basic particle filter methods popularly used — Bootstrap Filter and Auxiliary Particle Filter — in estimating the parameters in infectious disease progression models which are non-linear in nature. We propose a posterior particle filter algorithm and two single statistic posterior particle filter algorithms to enhance handling outliers in data. The posterior particle filter algorithm and the two single statistic posterior particle filter algorithms are shown to out-perform the traditional bootstrap and auxiliary particle filters in terms of accurately and consistently

estimating the parameters in compartmental SIR models. Besides, we proposed a re-sampling algorithm and compare it with the current popularly used re-sampling algorithm to show the importance of the re-sampling algorithm in helping improving the consistency of the particle filters.

Dengue is currently diagnosed using test algorithm determined by number of days after illness onset which cause the challenge of prevalence estimation as the sensitivity and specificity level of patients varies with different RNA and antibody level. In Chapter 3, we tried to address the challenge of adjusting the estimated prevalence and propose the way of estimating corresponding confidence interval incorporating the individual level sensitivity and specificity. We compared sensitivity, specificity for individual level benefits and average estimation errors and precision for surveillance purpose of both using single test and possible combination of multiple tests. Prevalence estimation adjustment can correct all test combinations. Using immunoassays targeting DENV nonstructural protein (NS1), the combination the NS1 and and IgM-capture immunoassays (ELISA) and the combination of NS1 and real-time reverse transcription polymerase chain reaction (RT-PCR) can statistically significant improving sensitivity of the tests without sacrificing the specificity and narrowing the confidence interval of prevalence estimation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Uncertainty Analysis of Species Distribution Models

1.1 Introduction

Species distribution models [48, 42, 66] are commonly used to predict the geographic distributions of animals or plants species. They are applied in species conservation [25], ecology [3], and other fields. Some SDMs, like the maximum entropy model, are used to predict the probability for the species being present. Others, like Poisson point process models, are used to model the intensity of the species per unit area.

Quantifying the uncertainty of maximum entropy models can help biologists allocate sampling efforts more efficiently. For places with the same probability estimate, different uncertainty estimates can help differentiate the need for further sampling effort. It may be possible to lower uncertainties in the estimates by choosing sampling locations carefully. However, the independence between sample units need to be guaranteed to maintain the independence assumptions underlying a maximum entropy model. Quantifying the

Chen, Xi, Nedialko B. Dimitrov, and Lauren Ancel Meyers. "Uncertainty analysis of species distribution models." PloS one 14.5 (2019): e0214190. My roles: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing original draft, review and editing. [13]

uncertainty also helps biologists have an idea of the amount of data sufficient to estimate probabilities across the geographic area. Knowledge of uncertainty can help answer questions such as: What is the benefit of collecting an additional 1000 presence only data points? What are low and high scenarios for the output estimates?

Unfortunately, most SDM methodology focuses on using point estimates. Point estimation involves using a single value for estimating target population parameters from sample data. However, the estimations are usually not equal to the target population parameters exactly, and so the accuracy of the estimations is important. A well accepted method of describing the uncertainty of the estimations is to look at their variance. With the variance of the estimates, one can compute confidence intervals, an interval that contains the true parameter with a certain confidence[44]. With current practice, SDMs only produce point estimates for the predicted probability or intensity at all species locations and background points without any corresponding uncertainty estimates at these locations.

To address this lack of uncertainty quantification in SDM, one must refer both to the SDM methodology and statistical methodology in quantifying uncertainty of point estimates. One of the most popular methods for SDM is the maximum entropy model. The conventional maximum entropy model was first formulated by Jaynes in 1957 [31] based on Shannon's measure of entropy [55] (see details in [32]). MAXENT incorporating the effect of actual occurrence data, became popular among biologists in modeling species distribution

with the contribution of MaxEnt software [48, 47, 19]. The mathematical equivalence of MAXENT, model used in MaxEnt software, and Poisson point process models (Poisson PPMS has been shown in [51]. Poisson PPMS may be fitted in the 'spatstat' package in R, which provides a way of assessing model uncertainty by providing standard error estimates [35]. To quantify uncertainty in point estimates, bootstrap methods are popular. Bootstrap uses computer-intensive simulation to calculate standard deviations of the estimated parameters, and is broadly applied in the biology field [18, 14, 39, 20]. In this chapter, we adopt the maximum entropy method and compare the analytical expression of the standard deviation with the standard deviation calculated through bootstrap method and Poisson point process model (PPM) approach.

In this chapter we consider quantifying the uncertainty in SDM. We focus specifically on the maximum entropy SDM methodology. A significant reason for the popularity of the maximum entropy methodology is its applicability to presence-only data with least assumptions [46]. For traditional statistical estimation methods like regression, both of the presence and absence of the species are required. However, in real cases, biologists often only know the places a species has been observed, while lacking information about absences of species.

The main contribution is analytically deriving an expression of the standard deviation of the target species distribution probabilities and comparing the results with bootstrap methods and standard deviation calculated

through Poisson PPM approach. We show that the three methods generate comparatively the same results and our analytic model uncertainty calculation procedure is dramatically faster than the boostrap method and more proper comparing to Poisson PPM without independence assumption and provided a direct result to maximum entropy model.

1.2 Materials and Methods

1.2.1 Maximum Entropy Model

Consider a region with geographic divisions given by $X = \{x_1, x_2, ..., x_n\}$. Suppose some species lives in the region, and the fraction of the species that lives in division i is p_i. A basic goal in SDM is to reconstruct the geographic distribution $P = \{p_1, p_2, ..., p_n\}$. To do this, we have some species occurrence data $O = \{o_1, o_2, ..., o_n\}$, where each o_i specifies the number of times the species has occurred in division i. The occurrence data can be viewed as a sample from the distribution P. In addition we had k layers of environmental data for the region described by features $f_j(X)$ for $j = 1, \ldots, k$. For example, one such function could be the average elevation in each geographic division.

Jaynes' maximum entropy model attempts to reconstruct P. Let $\hat{P} = \{\hat{p}_1, \hat{p}_2, ..., \hat{p}_n\}$ be the reconstructed density. Let $\hat{E}(f_j(X)) = (\sum_{i=1}^{n} o_i f_j(x_i))/(\sum_{i=1}^{n} o_i)$ be the empirical estimate of $E_P(f_j(X)) = \sum_{i=1}^{n} p_i f_j(x_i)$ given by the occurrence data O. Jaynes' maximum entropy model attempts to reconstruct P through an optimization problem. The optimization uses Shannon's measure of entropy as the objective (A.1a), subject to the moment constraints (A.1b).

Constraints (A.1c) and (A.1d) ensure that the optimal solution for the optimization is a probability distribution [32]. A mathematical formulation of the maximum entropy problem is

$$\max_{p_i} \quad -\sum_{i=1}^{n} p_i \log p_i \tag{1.1}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} p_i f_j(x_i) = \hat{E}(f_j(X)) \quad j = 1, \ldots, k \tag{1.2}$$

$$\sum_{i=1}^{n} p_i = 1 \tag{1.3}$$

$$p_i \geq 0 \qquad i = 1, \ldots, n \tag{1.4}$$

1.2.2 Bootstrap Method

Table 1.1 describes the bootstrap method for estimating the uncertainty of the estimate resulting from maximum entropy. The core of this bootstrap procedure is thinking of the distribution P as parameterized by the values it assigns to each geographic division. The procedure starts by estimating the parameters once, yielding a probability distribution. Then, it samples the data from that estimated distribution to construct several new estimates.

1.2.3 Analytic Deduction of Uncertainty

In this section, we demonstrate the basic idea of the analytical method for quantifying uncertainty in maximum entropy. The data $O = \{o_1, o_2, ...o_n\}$ follow a multinomial distribution with unknown parameters P. A maximum likelihood estimator for P follows a certain multivariate normal distribution as the number of samples grow large. The maximum entropy model can be

Table 1.1: Bootstrap Method

Algorithm — Bootstrapping	
1	**function** Bootstrapping (N)
2	$\hat{P} = M(O)$
3	**For** $i = 1 : N$ **do**
4	$\hat{O}_n = B(\hat{P}, m)$
5	$P' = M(\hat{O}_n)$
6	Record P'
7	Return $SD(P', N)$
N	Repeat the procedure N times
$O = \{o_1, o_2, ..., o_n\}$	Original occurrence data
$M(O_n)$	Fit a maximum entropy model given a set of species occurrence data $O_n = \{o_1, o_2, ..., o_n\}$ and return probability density estimation $\hat{p}$
$\hat{P}$ and P'	A reconstructed density over the geographic region
$B(\hat{p}, m)$	Sample m occurrence data following probability density $\hat{p}$, where $m = \sum_{i=1}^{n} o_i$
$\hat{O}_n = \{\hat{o}_1^n, \hat{o}_2^n, ..., \hat{o}_n^n\}$	The n^{th} new sampled occurrence data with $\sum_{i=1}^{n} \hat{o}_i^n = \sum_{i=1}^{n} o_i$
$SD(P', N)$	Calculate standard deviation of the set of $P's$

viewed as a function mapping this estimator to $\mathcal{R}^n$. The input is the empirical expectations, $\hat{E}(f_j(X))$, derived from the observation data, $O = \{o_1, o_2, ...o_n\}$. The output is the estimate of the probability distribution over geographic regions, $P = \{p_1, p_2, ..., p_n\}$. The analytical method of quantifying uncertainty describes how the output, P, changes as the input, O, changes. This is essentially a quantification of the way the optimization mapping warps the data input space, to the output space. We show the detailed deduction of the analytic method for uncertainty in the Appendix A.

For brevity, let $a_j = \hat{E}(f_j(X))$ and the vector of a_j can be expressed as $A = (a_1, a_2, ...a_k)^T$. Let $g(A)$ denote the maximum entropy optimization, model (A.1a),(A.1b),(A.1c),(A.1d), as a function from $\mathcal{R}^k$ to $\mathcal{R}^n$. In other words, the function takes as input the vector A with j^{th} entry specified by a_j, specifying right hand sides of the equality constraints $\hat{E}(f_j(X))$, and outputs a probability estimate across the geographic region P. We would like to understand the uncertainty in the output $g(A)$ as a function of the uncertainty of the input A. This can be done following steps similar to those in the delta method [2, p.75].

To understand the uncertainty in the output $g(A)$, we begin by writing a first order Taylor expansion of g around $E(A)$

$$\begin{aligned} g(A) &\approx g(E(A)) + \nabla g(E(A)) \cdot [A - E(A)] \\ &\approx g(\mathbf{F} \cdot \hat{P}) + \nabla g(\mathbf{F} \cdot \hat{P}) \cdot [A - E(A)], \end{aligned} \tag{1.5}$$

where $\mathbf{F}$ is $k \times n$ matrix of k features with entry (i, j) specified by $f_i(x_j)$ and $\nabla g(\cdot)$ is an $n \times k$ matrix of partial derivatives, with entry (i, j) specified by $\frac{\partial p_i}{\partial a_j}$. If we can compute an expression for these partial derivatives, then everything on the right hand side above is constant, except $[A - E(A)]$ whose distribution we know because we know the distribution of A. $g(A)$ is an affine transformation of $[A - E(A)]$, and can be approximated as

$$g(A) \sim \text{Normal}(g(\mathbf{F} \cdot \hat{P}), \nabla g \cdot \frac{\mathbf{F} \cdot \mathbf{\Sigma} \cdot \mathbf{F}^T}{m} \cdot (\nabla g)^T), \tag{1.6}$$

where $\mathbf{\Sigma}$ is proportional to the covariance matrix of $\hat{P}$ with entry (i, j) specified by $-\hat{p}_i\hat{p}_j$ for $i \neq j$, and entry (i, i) specified by $\hat{p}_i(1 - \hat{p}_i)$.

We express the $\frac{\partial p_i}{\partial a_j}$ as (Detailed deduction shown in S1 Appendix)

$$\frac{\partial p_i}{\partial a_j} = \sum_{r=1}^{k} p_i(a_r - f_r(x_i))((-\mathbf{\Psi})^{-1})_{rj}, \tag{1.7}$$

where $\mathbf{\Psi}_{rj} = cov_P(f_r, f_j)$ is the covariance matrix of features with respect to the maximum entropy model results, and f_j denotes the j^{th} feature in constraint (A.1b). We denote the inverse covariance matrix as $\mathbf{\Psi}^{-1}$ and refer to its (r, j)th entry as $(\mathbf{\Psi}^{-1})_{rj}$.

To summarize, one can compute analytical estimates of the uncertainty as follows:

1. Gather data for $f_r(\cdot)$ and the right-hand sides of constraints (A.1b), a_r.
2. Solve the maximum entropy model to get a vector of P of probabilities p_i.
3. Compute the matrix $-cov_P(f_r, f_j)$, using the features and the vector P.
4. Compute the derivates $\frac{\partial p_i}{\partial a_j}$ using (1.7), giving the matrix ∇g.
5. The covariance of the output P can then be estimated as $\nabla g \cdot \frac{\mathbf{F}\cdot\mathbf{\Sigma}\cdot\mathbf{F}^T}{m} \cdot (\nabla g)^T$, following equation (1.5).

1.3 Results

We demonstrate the applications of the analytical expression of the uncertainty through two examples, Dengue virus and Aedes Aegypti mosquito, and compare the analytical results with the uncertainty calculated using the bootstrap method and Poission PPM approach. The analytic method results aligned well with bootstrap method results, but Poisson PPM approach gave much larger standard deviations. We only show the results and comparison of analytic and bootstrap below but include results and comparison of Poisson PPM in Fig A.1. The resolution of the Dengue virus example is at county level while the resolution of the Aedes Aegypti mosquito is at 1 km^2 area level through Texas.

Dengue Importation Probabilities. Dengue virus is often imported into Texas from endemic counties. We aim to estimate the probability that the next importation case will happen in each county of Texas. Historical case import data, $O = \{o_1, o_2, ...o_n\}$ with n equal to 254 counties in Texas, present empirical samples from this distribution. Each o_i counts the number of imports in county i. We am also given features $f_j(X) \in R^{1\times 254}$ for $j = 1, ..., 10$ that represent socio-economic, demographic, and environmental features selected for all 254 counties across the Texas counties. This completely defines the inputs necessary for a maximum entropy model.

Specifically, we use ten years, 2002 to 2012, of Dengue importation data into Texas received from the Texas Department of State Health Services. The

features $f_j(X)$ represent features listed in Table 1.2. The ten final features were selected through a series feature selection procedures, including representative variable selections and most predictive variable selections, which demonstrated in [7]. We estimate the standard deviation using the bootstrap method, Poission PPM approach and the analytic method. The results are presented in Fig 1.1.

Table 1.2: Features for modeling Dengue importation

Features
Population of Educational Attainment with Bachelor's degree
Minimum Temperature of Coldest Month
Percentage of Using Public Transportation to Work
Population of Educational Attainment in some college(no degree)
Population of Walked to Work
Population of Commuting to Work with Other Means
Population of Educational Attainment less than 9th grade
Percentage with Graduate or professional degree
Percentage of Walked to Work
Average Artificial Surface (Percentage)

Ten features included in maximum entropy model. The data for these features is derived at a county level from the 2009-2013 American Community Survey 5−year estimates [61] and WorldClim Database [28]

Fig 1.1a shows the point estimates for the import probability estimated from maximum entropy model and Fig 1.1b represents standard deviation estimates from the bootstrap method and analytic method of maximum entropy model, respectively. Many Texas counties have never had imported Dengue cases over the past ten years, and their estimates are close to zero. We map the

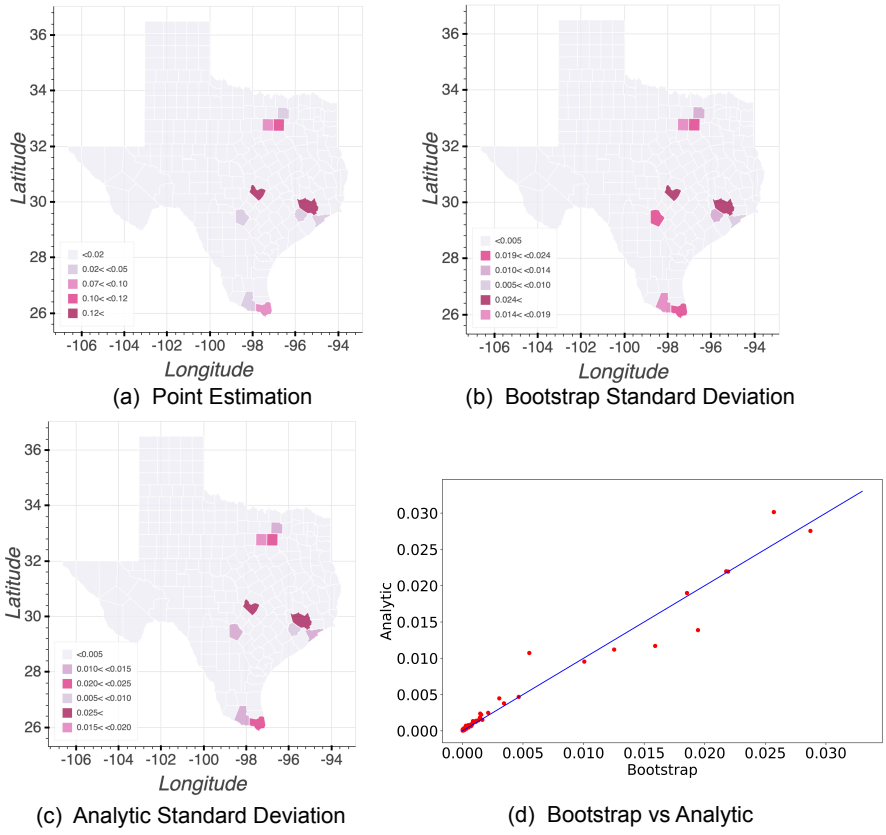


Figure 1.1: **Standard deviation comparison for Dengue importation probability.** (a) Figure shows the point estimates for the import probability $\hat{p}_i$. (b) Figure visually plots the bootstrap standard deviation estimates for p_i across Texas counties. (c) Figure visually plots the analytic standard deviation estimates for p_i across Texas counties. (d) Figure plots the standard deviations of bootstrap vs. analytic and shows a strong equivalence between the two. Each red dot represent the estimations for one county

standard deviation of the estimates p_i of each county in Fig 1.1b and Fig 1.1c with a darker color indicating a higher standard deviation level. For bootstrap method, we did 2000 bootstrap runs and took 22403.34 seconds in total. The running time of the analytic method, using optimized matrix operations as described in the Appendix A, is dramatically faster than the bootstrap method and takes 0.0016 seconds in total.

Fig 1.1d shows the standard deviation resulting from the bootstrap against the standard deviation resulting from the analytic method. Each red dot represent a county. It also depicts a regression line between the two results $s_a = 0.98 s_b$ with $R^2 = 0.972$, where s_a and s_b stand for the standard deviation estimates from the analytic and the bootstrap methods, respectively. Regression results show a linear relationship between the standard deviation calculated from analytic expression and bootstrap method with parameter approximate 1. Both bootstrap method and analytic method generally indicate larger standard deviation for counties with larger point estimates.

Aedes Aegypti Habitat The Aedes aegypti mosquito is the primary transmission vector of dengue, chikungunya, and zika viruses. We aim to estimate the relative probability distribution of Aedes aegypti in Texas. Historical presence data $O = \{o_1, o_2, ...o_n\}$, with n equal to the number of 1km grid squares in Texas, present empirical samples from this distribution. Each o_i is either 0, if there is no presence data for this square, or 1 if there is presence data. The features $f_j(X)$ represent environmental data for each 1 km^2 area across

the Texas.

Specifically, we use 121 locations, within Texas, of Aedes aegypti presence data found from previous studies [64, 41, 40, 24, 6, 33, 4, 57], DSHS. The environmental features $f_j(X)$, found from WorldClim Database [28], are listed in Table 1.3.

We aim to analyze the standard deviation of the estimates $\hat{p}_i$ for each 1km square. We estimate this standard deviation using both the bootstrap method and the analytic method. The results are presented in Fig 2.2.

Table 1.3: Features for modeling Aedes aegypti existence

Features
artificial surfaces
population count
temperature seasonality
elevation
precipitation seasonality
minimum temperature of coldest month
mean diurnal range

Seven features, found from WorldClim Database [28], included in maximum entropy model

We present the point estimates of the distribution of the Aedes aegypti mosquito in Fig 2.2a. Aedes aegypti primarily feeds on humans and is found in urban areas, which results in higher probability estimates in those areas. The areas of concentration of Aedes aegypti in Texas tend to be population centers like Houston, Dallas, San Antonio, Austin, El Paso, and McAllen.

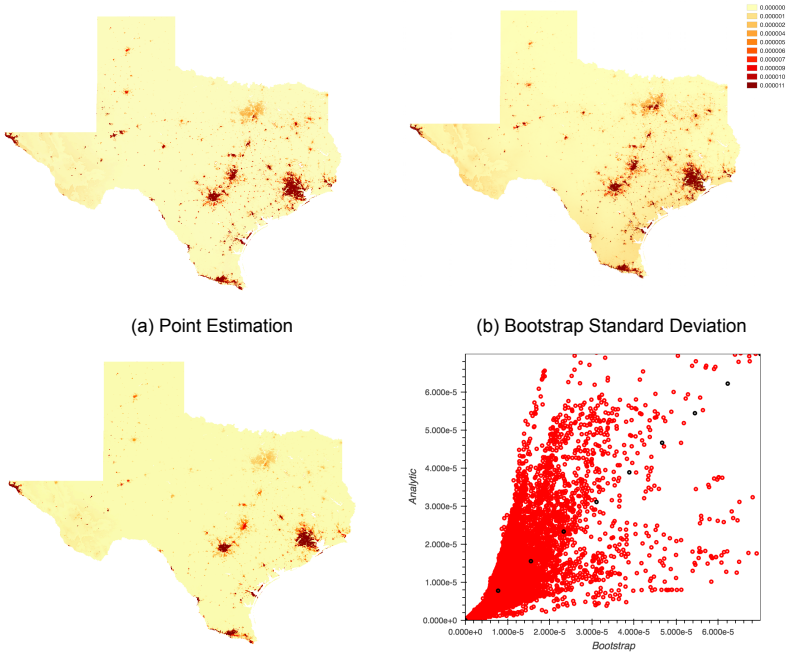


Figure 1.2: **Standard deviation comparison for Aedes aegypti.** (a) Figure presents the point estimates p_i. (b) Figure shows standard deviation calculated using bootstrap method. (c) Figure shows standard deviation calculated using analytic method. (d) Figure shows the standard deviation comparison between analytic method and bootstrap method.

Fig 2.2b and fig 2.2c plots the standard deviation of the estimates p_i of each grid using the bootstrap method and the analytic method. This can give a practitioner a good sense of the standard deviation in the estimates. In applying the analytic method, one could use as input the empirical distribution or a Laplace smoothed estimator [53] to smooth the empirical probability to be non-zero. The analytic method gives slightly higher uncertainty estimates than bootstrap as shown in Fig 2.2c. Each red dot represent the standard deviation estimates for each grid using the bootstrap and the analytic method respectively. The black dot shows the diagonal line when two methods aligned well. When we applied a Laplace smoothing of 0.0001, we have the relationship $s_a = 1.0744 s_b$ with $R^2 = 0.802$, where s_a and s_b stand for the standard deviation estimates from the analytic and the bootstrap methods, respectively. We map the standard deviation of the estimates p_i of each 1 km^2 using the analytic method and Laplace smoothing of 0.0001 in Fig 2.2b. The bootstrap result and analytic result can be visually compared through Fig2.2 c.

We did 2000 bootstrap runs and took 30400 seconds in total. The running time for the analytic method is 6516 seconds, which is much faster than bootstrap method. As we calculated the relative probability for Aedes aegypti for a $1km^2$ square grid, we have $933,680$ grid cells in total. Computing the covariance of the output would require matrix multiplications for matrices of size 933680×933680, which can cause out-of-memory errors. We introduce a faster method of calculating the variance of each square grid in Appendix A.

1.4 Discussion

The maximum entropy model can give a point estimation of the unknown species distribution within predefined grids using presence-only data with possible influential features like environmental factors, demographic factors, social economic factors, etc. However, uncertainties come from both the model and the sample data. Some possible sources of this uncertainty are:

- The true expectation of all features $f_j(X)$ are unknown and estimated using the presence-only data.
- Species distribution data are not collected at random, but based on prior knowledge of the biologists. For example, all samples may be observed within pre-selected locations.
- Having only a few presence points relative to the size of the grid can lead to unstable models.
- The features $f_j(X)$ used within the model may be inaccurate or vary dramatically over time. So, it is unclear whether the presence only data collected is appropriate for use with the given features.

In the maximum entropy model, the output probabilities are dependent on the features $f_j(X)$. A flat $f_j(X)$ can only produce flat output probabilities. One may want to know how will the output probabilities change when the feature values change? Uncertainty quantification may help identify the features that most reduce uncertainty in a maximum entropy model.

The bootstrap method is a well accepted method of quantifying uncertainty. However, the running time of the bootstrap can be very long. In the dengue example, bootstrap method took more than 22000 seconds to generate a comparable uncertainty estimate of analytic method while the analytic method just took 0.0016 seconds in total. The analytic method uses more memory compared to the bootstrap method. In the Aedes aegypti example, the analytic method took only 20 percent time of running bootstrap method. However, code optimization, and element-wise matrix multiplications can significantly increase the speed of the analytic method compared to the bootstrap method. A method for increasing programming speed are shown in Appendix A. Furthermore, the analytic method is able to approximate covariances in the output – whereas this can be quite difficult for the bootstrap method if Weonly use a small number of samples.

The Poisson PPM approach proved to be equivalent to MAXENT providing an alternative approach of estimating the uncertainty. However, the hidden independence assumption of species appearance locations can affect the performance of the model which gives much larger estimated uncertainty when assumptions violated.

Chapter 2

Comparison of Particle Filter Methods for the Estimation of Reproduction Number of Influenza Epidemics

2.1 Introduction

The influenza, which also called Flu, is a contagious respiratory disease caused by the influenza viruses infecting nose, throat and lungs with severity ranks from mild to severe [10]. In 2003, in the US alone, about 610K loss of life-years, more than 3 million hospitalized days, more than 31 million outpatient visits and annual medical cost of 10.4 billion are caused by influenza epidemics [43].

Basic reproduction number, denoted by R_0, is the expected number of new infections created by an infected individual under the most favorable conditions for transmission. Any infectious disease can become epidemic in the host population if $R_0 > 1$ [15]. In this chapter, we focus on studying the methods of estimating the quantity of R_0 which can address the fundamental question of "Which influenza become epidemics?" We use SIR model, Susceptible-Infectious-Recovered model, to explore the performance of particle filters in recovering true parameters.

There are many published studies on disease progression parameter estimation. Some of them using multi-linear regressions [54, 22], while some are based on alternate statistical techniques like the method of analogues [63]. A problem associated with the current infectious disease modeling is the disability of incorporating the non-linear mechanism of the disease progression model. Particle filter can overcome this limitation.

Filtering method is the method of estimating the states, including parameters and hidden variables, of a system base on available observations [62]. Particle filter performs sequential Monte Carlo (SMC) estimation based on point mass (called "particle") representation of probability densities [52]. The first SMC idea was introduced in statistics in 1950s [27] with sequential importance sampling considered only and became much more popular in applications after re-sampling step was incorporated [23]. The re-sampling step is used to eliminate the particles with low importance weight and keep the particles with high importance weight [23]. Particle filter depends on the importance sampling that requires a well designed proposal distribution, which is called importance density and used to sample particles, that can approximate the posterior distribution reasonably well [62].

First, we introduced a modified Bootstrap Filter method which improves the performance of the Bootstrap Method in terms of accuracy, consistency and the ability of handling outliers in observation data. Second, we introduced two modified auxiliary particle filter methods which outperform the original auxiliary particle filter described in [1]. Third, we showed that the

historical data can be efficiently used in preventing particles moving toward wrong directions with the effects of the outliers in historical observation data. We also showed that it is worth to collect more information about infectious pandemic as they can be very useful in terms of helping particle filter methods finding a more accurate and consistent estimation of the true parameters. Fourth, we studied the possibility of using particle filter methods to find the time for the outbreak to start which is useful for helping public organizations react quickly and plan infectious pandemic strategies more efficiently. Fivth, we showed that the re-sampling procedure is crucial for particle filter methods to perform consistently, and importance sampling procedure influence the accuracy of the particle filter methods in finding the true parameters. Finally, we showed that the number of the particles in each particle filter method can influence the consistency of the method. More particles help improve the performance of finding true β and γ parameter in compartmental SIR model, but had no significant effect in finding the true outbreak time.

2.2 Materials and Methods

2.2.1 Simulation Data

In the following sections, we denote the number of patients that go to the doctor as *Doctor Visit Data*, the length of the period for which a patient can transmit the disease as *Infectious Period Data*, and the number of people secondary infections from a single infected individual as *Contact Tracing Data*. We assume the total population is $100,000$. First, we simulated the number of

people infected (denoted as I in SIR model) by passing the true β and γ value into the compartmental SIR model. We assumed 50% of people infected will see a doctor, and simulated the number of patients that go to the doctor using a binomial distribution $Bin(I, 0.5)$. We generated the infectious period data for each patient based on an exponential distribution, i.e e^{γ}, with expected length equal to $\frac{1}{\gamma}$. For contact tracing data, we assumed that there are 100 people in contact with an infected individual. The number new infections follows a binomial distribution, i.e $Bin(100, \beta\frac{S}{S+I+R})$.

2.2.2 Compartmental SIR Model

The compartmental SIR model used in this paper is a simple and widely used differential equation model of infectious disease [34]. Assuming the total population is N and separated into three parts: (1) **S**, stands for *susceptible*, representing the portion of the people who are never infected by the disease but susceptible of being infected (2) **I**, stands for *infected*, representing the portion of people currently infected and (3) **R**, stands for *resistant*, representing the portion of people who do not have disease, cannot infect others and cannot be infected. [15]

The compartmental SIR model equations are:

$$\frac{dS(t)}{dt} = -\beta \cdot S(t) \cdot I(t) \tag{2.1a}$$

$$\frac{dI(t)}{dt} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t) \tag{2.1b}$$

$$\frac{dR(t)}{dt} = \gamma \cdot I(t). \tag{2.1c}$$

For the SIR model, $R_0 = \beta/\gamma$, with β to be the transmission rate and γ parameter with its reciprocal $\frac{1}{\gamma}$ determines the average infectious period [34]. As SIR model can be fully characterized by the parameters β and γ, we can easily build up the estimation for the fraction of population being susceptible, infected and recovered, which helps health organization determine the severity of the epidemics and response properly and quickly.

2.2.3 Particle Filters

We describe the basic idea of particle filter methods for influenza progression parameters based on [49] as follows. Assume we have reported time series influenza observations $z_t, t = 1, 2, ..., n$ conditionally independent given unobserved state sequence $x_t, t = 0, 1, ..., n$ which itself is a Markov chain. Each x_t describes the estimation of unknown parameters of the disease – β, γ at time t and the initial conditions of the differential equations in eq.2.1. The number of people that are susceptible, infected, and resistant at time period t determined by the unknown parameters. We used a particle filter to estimate the parameter values based on a time series of observations Z_t. These observations represent the doctor visit data, infectious period data, and contact tracing data. The particle filter works by estimating the probability density function $f(x_t|z_1, z_2, ...z_t) = f(x_t|z_{1:t})$. Filter methods, use two basic steps: 1. propagate the current states to future states via prediction density, LHS of (2.2a). 2.update the filtering density, LHS of (2.2b), by adopting new

available data through Bayes' rule as follows,

$$f(x_{t+1}|z_{1:t}) = \int f(x_{t+1}|x_t)dF(x_t|z_{1:t}) \tag{2.2a}$$

$$f(x_{t+1}|z_{1:t+1}) = \frac{f(z_{t+1}|x_{t+1})f(x_{t+1}|z_{1:t})}{f(z_{t+1}|z_t)}. \tag{2.2b}$$

$$f(z_{t+1}|z_t) = \int f(x_{t+1}|z_{t+1})dF(x_{t+1}|z_{1:t}). \tag{2.2c}$$

Particle filter methods, in contrast with filter methods, recursively approximate the random variable $x_t|z_{1:t}$ using particles $x_t^1, x_t^2, ...x_t^M$ associated with probability mass density $w_t^1, w_t^2, ...w_t^M$. The prediction density and filtering density, LHS of (2.2a) and (2.2b), then can be approximated as

$$f(x_{t+1}|z_{1:t}) = \sum_{i=1}^{M} f(x_{t+1}|x_t^i)w_t^i \tag{2.3a}$$

$$f(x_{t+1}|z_{1:t+1}) = f(z_{t+1}|x_{t+1}) \sum_{i=1}^{M} f(x_{t+1}|x_t^i)w_t^i. \tag{2.3b}$$

Particle filter methods have many advantages. Posterior distributions of parameters conditional on the reported influenza cases cannot be expressed analytically. With particle filters, these posterior distributions can be represented by a set of *particles* $x_t^1, x_t^2, ...x_t^M$ associated with a probability mass density $w_t^1, w_t^2, ...w_t^M$ [1]. In general, the unknown parameters may vary over time with a *transition density* $f(x_{t+1}|x_t)$. Moreover, the likelihood function $f(z_t|x_t)$ expressing the probability of an observations given disease parameter values is non-linear. Particle filter methods can easily incorporate the non-linearities and transition densities, which makes them superior to many popular algorithms like Kalman Filter and extended Kalman filter.

2.2.3.1 Bootstrap Filter

Bootstrap filter were one of the earliest implementations of particle filters, proposed in [23] and motivated by sampling importance re-sampling (SIR) proposed in [56]. SIR for particle filters refers to a method quite separate from the SIR differential equation model of disease progression. Details of the particle filter methods can be found in [16], named SIS/Re-sampling Monte Carlo filter algorithm, or as the algorithm 4, named as SIR Particle Filter, in [1]. For estimating disease parameters, the bootstrap filter propagates and updates the *particles* $\{x_t\}_{i=1}^{M}$ given new data z_{t+1} as follows. Firstly, $\hat{x}_{t+1}^i$ are propagated from x_t^i based on a transition density $f(x_{t+1}^i|x_t^i)$. Then, weights for $\hat{x}_{t+1}^i$ are calculated proportional to the likelihood of the most recent observations conditional on the $\hat{x}_{t+1}^i$, i.e $w_{t+1}^i \propto w_t^i f(z_{t+1}|\hat{x}_{t+1}^i)$. Finally, new particles x_{t+1}^i at time $t+1$ are sampled from $\hat{x}_{t+1}^i$ with replacement based on probabilities specified by w_{t+1}^i.

In the bootstrap filter method, $\hat{x}_{t+1}^i$ were propagated without considering the most recent observation data z_{t+1}. When we implement the bootstrap filter algorithm, we find that particles are easily propagated to wrong directions, and whenever this situation happens, the particles can be stuck in an area where newly observed data has zero probability. In the result section, we show that the bootstrap filter fails to work for some cases.

2.2.3.2 Posterior Particle Filter

To handle the inefficiency of the bootstrap filter not considering the most recent observation, we propose the posterior particle filter (PPF), as an ad-how algorithm for approximate posterior updating. This method updates the weights incorporating the most recent observation data while also using the posterior distribution at time t as the prior for the updating at time $t+1$. In the re-sampling procedure, particles with low likelihood for the observation data will be re-sampled using Algorithm 4 in table 2.4 and particles with high likelihood will be kept. For all the new particles $\{x_{t+1}\}_{i=1}^{M}$, their weights will be calculated by applying all the historical data to minimize the effects of the outliers in the observation data. Details of the PPF method are showed in table 2.1.

Table 2.1: Algorithm 1: Posterior Particle Filter

Algorithm 1: Posterior Particle Filter	
$\{x_{t+1}^i, w_{t+1}^i\}_{i=1}^M = PPF\{x_t^i, w_t^i, z_{t+1}, z_{1:t}\}_{i=1}^M$	
1	**For** $i = 1 : M$ **do**
	update $w_{t+1}^i = f(z_{t+1}\|x_t^i)w_t^i$
2	**Normalize** $w_{t+1}^i = w_{t+1}^i / \sum_{i=1}^M w_{t+1}^i$
3	**Resample** using Algorithm 4 in table 2.4
	$\{x_{t+1}^i, i_k, 1/M\}_{i=1}^M = Resample\{x_t^i, w_{t+1}^i\}_{i=1}^M$
	set $w_0^i = 1/M$
4	**For** $i = 1 : M$ **do**
	For $\tau = 0 : t$ **do**
	update $w_{\tau+1}^i = f(z_{\tau+1}^i\|x_{t+1}^i)w_\tau^i$

2.2.3.3 Auxiliary Particle Filters

Auxiliary Particle Filters (APF), first proposed in [49], was developed to overcome the weakness of other particle filters based on SIR which perform poorly when there are outliers, and without greatly slowing the running time of the filter [49]. The idea of APF is to sample from the joint distribution of the states and the auxiliary variable conditional on the available observations,

$$f(x_{t+1}, \mu^i_{t+1}|z_{1:t+1}) \propto f(z_{t+1}|x_{t+1}) f(x_{t+1}|x^i_t) w^i_t,$$

and then discard the auxiliary variable. Weights for the new states can be constructed based on equation (72) in [1]. In equation (72), the auxiliary variable μ^i_{t+1} is some characterizations associated with $f(x^i_{t+1}|x^i_t)$, for example the mean, the mode, a draw or any other likely value. In this paper, we sample μ^i_{t+1} from a transition density $f(x^i_{t+1}|x^i_t)$.

The algorithm we used to implement APF is the algorithm 5 in [1]. First, we sampled the auxiliary variables for each particle x^i_t from the transition density $\mu^i_{t+1} \sim f(x^i_{t+1}|x^i_t)$. Then, we calculated the re-sampling weights for each particle x^i_t as $\lambda^i_{t+1} \propto f(z_{t+1}|\mu^i_{t+1})\lambda^i_t$. After normalizing $\lambda^i_{t+1} = \lambda^i_{t+1} / \sum_{i=1}^M \lambda^i_{t+1}$, for $i = 1, 2, ..., M$, we re-sampled from current particles x^i_t as follows. First, for each particle i, we selected a particle $x^{i^j}_t$ based on the normalized probabilities λ^i_{t+1} using algorithm 2 in [1]. Then, we propagated the selected particles $x^{i^j}_t$ based on the transition density $f(x^i_{t+1}|x^{i^j}_t)$. The probability mass density for new particles x^i_{t+1} were then calculated based on equation (72) in [1].

2.2.3.4 Single Statistic Posterior Particle Filters

Single Statistic Posterior Particle Filters 1 To enhance the ability of APF in handling outliers and data variations, we propose two new algorithms and describe our first algorithm — single statistic posterior particle filters 1 (SSPPF1) as Algorithm 2 in table 2.2. We named the method *single statistic posterior particle filter* as the auxiliary variable μ_{t+1}^i can be viewed as a single statistic of $x_{t+1}^i|x_t^i$ and we also use the posterior distribution at time t as the prior for the updating at time $t+1$. The major difference between the APF and SSPPF1 are the re-sampling procedures. We use a re-sampling algorithm described in Algorithm 4 in table 2.4 which makes the method only re-sample particles which give the new observation a low likelihood. In the results section, we show that our SSPPF1 performs much better in providing stable and accurate estimations.

Table 2.2: Algorithm 2: Single Statistical Posterior Particle Filters 1

Algorithm 2: Single Statistical Posterior Particle Filters 1	
	$\{x_{t+1}^i, w_{t+1}^i\}_{i=1}^M = SSPPF1\{x_t^i, w_t^i, z_{t+1}\}_{i=1}^M$
1	**For** $i = 1 : M$ **do**
	sample $\mu_{t+1}^i \sim f(x_{t+1}^i|x_t^i)$
	calculate $\lambda_{t+1}^i \propto f(z_{t+1}|\mu_{t+1}^i)\lambda_t^i$
2	**Normalize** $\lambda_{t+1}^i = \lambda_{t+1}^i / \sum_{i=1}^M \lambda_{t+1}^i$
3	**Resample** using Algorithm 4 in table 2.4
	$\{x_{t+1}^i, i_k, 1/M\}_{i=1}^M = Resample\{x_t^i, \lambda_{t+1}^i\}_{i=1}^M$
4	**For** $i = 1 : M$ **do**
	$w_{t+1}^i = f(z_{t+1}|x_{t+1}^i)/f(z_{t+1}|\mu_{t+1}^{i_k})$

Single Statistic Posterior Particle Filters 2 Historical observation data can be applied to reduce the influence of observations that are outliers, similarly to the posterior particle filter algorithm. In SSPPF2, we update the new weights of all particles by applying all the historical data as described in Algorithm 3 in table 2.3. In the results section, we show that Algorithm 3 outperforms all the other particle filter methods in this paper by providing the most stable and accurate estimations of the true parameter.

Table 2.3: Algorithm 3: Single Statistical Posterior Particle Filters 2

Algorithm 3: Single Statistical Posterior Particle Filters 2	
1	**For** $i = 1 : M$ **do**
	sample $\mu^i_{t+1} \sim f(x^i_{t+1}\|x^i_t)$
	calculate $\lambda^i_{t+1} \propto f(z_{t+1}\|\mu^i_{t+1})\lambda^i_t$
2	**Normalize** $\lambda^i_{t+1} = \lambda^i_{t+1} / \sum_{i=1}^M \lambda^i_{t+1}$
3	**Resample** using Algorithm 4 in table 2.4
	$\{x^i_{t+1}, i_k, 1/M\}_{i=1}^M = Resample\{x^i_t, \lambda^i_{t+1}\}_{i=1}^M$
4	**For** $i = 1 : M$ **do**
	$w^i_{t+1} = f(z_{t+1}\|x^i_{t+1}) / f(z_{t+1}\|\mu^{i_k}_{t+1})$
	set $w^i_0 = w^i_{t+1}$
5	**For** $\tau = 0 : t$ **do**
	update $w^i_{\tau+1} = f(z^i_{\tau+1}\|x^i_\tau) w^i_\tau$

Re-sampling Algorithm We introduced a re-sampling algorithm as Algorithm 4 in table 2.4 which keeps the particles with high conditional likelihood for the observations to occur and re-samples the particles with low conditional likelihood. The re-sampling algorithm is described in Table 2.4. The advantage of this re-sampling algorithm is the improvement of efficiency as we can

move towards true parameters faster. Also, this algorithm can efficiently help prevent the particles from getting stuck in an area where new observations have zero conditional probability.

Table 2.4: Algorithm 4: Re-sampling Algorithm

Algorithm 4: Re-sampling Algorithm	
	$\{x_{t+1}^i, i_k, p_{t+1}^i\}_{i=1}^M = RESAMPLE\{x_t^i, p_t^i\}_{i=1}^M$
1	**For** $i = 1 : M$ **do**
1.1	**If** $p_t^i < 1/M$
	sample index k base on probability p_t^i
	set $i_k = k$
	sample $x_{t+1}^{i_k} \sim f(x_{t+1}^{i_k} \mid x_t^{i_k})$
	set $x_{t+1}^i = x_{t+1}^{i_k}$
1.2	**Else**
	set $x_{t+1}^i = x_t^i$
	set $p_{t+1}^i = 1/M$
2	**Record** $\{x_{t+1}^i, i_k\}_{i=1}^M$

2.3 Results

2.3.1 SIR Model with Two Parameters

In this section, we compared the performance of the bootstrap filter, PPF, APF, SSPPF1 and SSPPF2 for estimating the parameters β and γ in a compartmental SIR model to answer two questions: "Which particle filter algorithm performs better?" and "Is it worth to collect more information?". We wanted to evaluate the performance of each particle filter algorithm based on the accuracy and the consistency of finding the true parameter values. We characterized each particle filter's performance through a confidence ellipse.

The key aspects of the confidence ellipse are how close its center is to the true parameter values, and the size of its volume. Algorithms whose ellipses have centers closer to the true values are more accurate, and algorithms whose ellipses have small volumes are more consistent.

For each particle filter method, we did 100 runs with each run contained 40 iterations based on 100 sets of simulated data which were generated using the data simulation method described in section *Simulation Data* by passing the true parameter values $\beta = 0.21$ and $\gamma = 0.07$, to get 100 parameter estimations. We plotted the standard deviation ellipse of these 100 estimations in 2 standard deviations for comparison. For each run in the simulation, we randomly generated data base on the true parameters to allow the existence of data variations, errors and outliers and test the ability of each particle filter in finding the true parameters. In each sub-figure, we plotted the true parameter value as a black dot for visual comparison.

Particle Filter Algorithms Comparison "Which particle filter algorithm performs better?" To answer the question, we compared the distance between the true parameter values and the center of the standard deviation ellipses of each particle filter algorithm. We defined a more accurate algorithm as its standard deviation ellipse centering closer to the true parameters. Also, we compared the size of the standard deviation ellipses of each particle filter algorithm and define the particle filter algorithm to be more consistent as having a smaller standard deviation ellipse.

Bootstrap algorithm performed the worst and APF performs better than it but worse than the others. In fig 2.1, no matter what information we included in the model, the bootstrap filter always performed worse than all the others by having a much larger standard deviation ellipse, which showed that observation data variations, errors and outliers can dramatically influence the accuracy and consistency of the Bootstrap Filter algorithm. APF performed better than Bootstrap Filter as the most recent observation are incorporated while calculating the re-sampling weight. But it performed worse than SSPPF1, SSPPF2 and PPF algorithm with a much larger standard deviation ellipse indicating a weaker ability of handling with data variations.

SSPPF1, SSPPF2 and PPF performed much better with a much smaller standard deviation ellipse centering at the true parameter values. SSPPF2 performed the best no matter what information were included in the model showing a great ability of handling data variation and outliers. In fig 2.1, with doctor visit data only or using both contact tracing data and doctor visit data, SSPPF1 and PPF performed similar to each other. With all information added, PPF could outperform SSPPF1, while with infectious period data and doctor visit data, SSPPF1 could outperform PPF. In general, SSPPF1 and PPF performed similar to each other.

Importance of Data Collection "Is it worth to collect more information?" To answer the question, we compared how the accuracy and consistency of each particle filters changed by having different data information. Generally

speaking, more data information can help particle filter algorithms provide more stable, consistent and accurate estimations.

Doctor visit data along was not sufficient for some particle filter algorithms. In fig 2.1, with doctor visit data only, the standard deviation ellipse generated using APF centered far away from the true parameter values indicating a failure of the algorithm recovering true parameters. Also, bootstrap algorithm gave very large ellipse indicating a highly unstable performance.

Adding more data can efficiently improve the performance of all the particle filter algorithms, but it is not true that the more data the better performance. Comparing either fig 2.1(c) and 2.1(e) with 2.1(a), or fig 2.1(d) and 2.1(f) with 2.1(b), adding more information can significantly improve the performance of all particle filter algorithms with all standard deviation ellipses shrinking dramatically and APF centering at the true parameters. However, if comparing fig 2.1(e) and 2.1(f) or fig 2.1(f) and 2.1(h), small marginal benefits showed up for bootstrap PF, APF, and SSPPF1 by having all three data sources instead of contact tracing and doctor visit data only.

Contact tracing information could be more useful in stabilizing the performance of the particle filter algorithms. Comparing fig 2.1(c) and 2.1(e) or fig 2.1(d) and 2.1(f), the standard deviation ellipses of all particle filter algorithms combined doctor visit data with contact tracing instead of infectious period were much smaller indicating a more consistent performance in handling data variation. Comparing fig 2.1(e) and 2.1(g) or fig 2.1(f) and 2.1(h), by having contact tracing data, adding infectious period data did not help in

shrinking the ellipses of bootstrap PF, APF, and SSPPF1.

2.3.2 Particle Filter for Three Parameters

In this section, we compared the performance of the bootstrap filter, PPF, APF, SSPPF1 and SSPPF2 in estimating three parameters — β and γ in compartmental SIR model and the time t for outbreaks to happen to answer a sequence of questions: "Can we estimate the outbreak time using particle filter methods?", "Which particle filter algorithm performs better with more unknown parameters?" and "Does it worth to collect more information?" Same as section *Particle Filter for Two Parameters*, we wanted to evaluate the performance of each particle filter algorithm base on the accuracy and the consistency of finding 3 true parameter values. We also did 100 runs with each run contained 40 iterations base on 100 sets of simulated data for each particle filter algorithm by passing the true parameter values $\beta = 0.21$, $\gamma = 0.07$ and $t = 6$, to get 100 parameter estimations. We calculated the volume of the standard deviation oval in table 2.5 for comparison and plotted the standard deviation ellipse of these 100 results in 2 standard deviations for comparison in fig 2.2.

Outbreak Time Estimation "Can we estimate the influenza outbreak time using particle filter methods?" Without any information for the outbreak time, it was not easy for the particle filter algorithms to perform good in predicting the outbreak time. The outbreak time estimations for all the particle filter

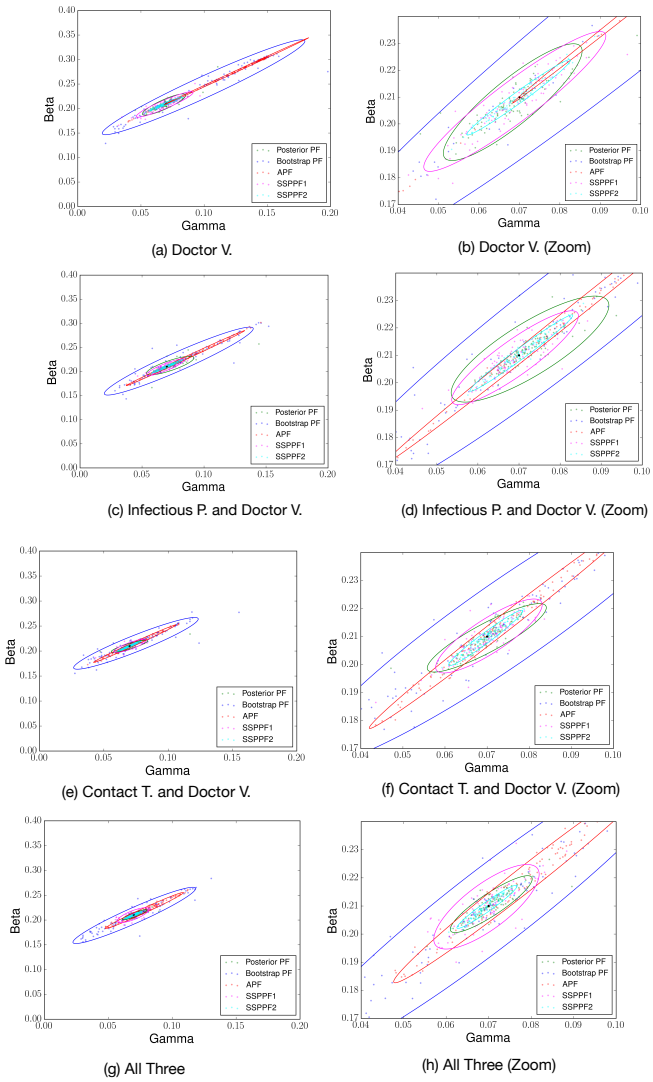


Figure 2.1: Fig (a) and (b) shows the standard deviation ellipses of 100 parameter estimations of each particle filter methods in finding the true parameters by having doctor visit data only. Fig (c) and (d) shows the standard deviation ellipses of 100 parameter estimations of each particle filters by having one more type of data — infectious period data. Fig (e) and (f) shows the standard deviation ellipses of 100 parameter estimations of each particle filters by having contact tracing data and doctor visit data. Fig (g) and (h) shows the standard deviation ellipses of 100 parameter estimations of each particle filters by having all three types of data.

algorithms skewed down to 0 with most of the estimations are smaller than the true outbreak time. Bootstrap method and APF tended to skewed far away while the other three algorithms performed better by giving a much closer standard deviation ellipse to the true outbreak time. PPF outperformed other algorithms by providing a flat oval close to the true outbreak time. SSPPF2 and SSPPF1 were able to give a relative good estimation about outbreak time as the standard deviation ellipses were relatively small and centering close to the true parameter values.

Particle Filter Algorithms Comparison "Which particle filter algorithm performs better with more unknown parameters?" With one more parameter to estimate, no matter what data information we have, the APF and bootstrap algorithm cannot functional well by centering far away from the true parameters, for either γ, β, or t. The ability of estimating the β and γ were lost by having one more parameter to estimate, as their standard deviation ellipses are centered away from the true β and γ. PPF, SSPPF1 and SSPPF2 performed relatively good as the ability of estimating the true β and γ values are retained and they also have a promising ability of estimating outbreak time. With one more parameter to estimate, SSPPF2 outperformed all other methods and modified bootstrap and SSPPF1 performed similar to each other.

Importance of Data Collection "Does it worth to collect more information?" Same as section *Particle Filter for Two Parameters*, it was worth

to collect more information for infectious disease progression parameter estimation. Contact tracing data tended to be more valuable comparing to the infectious period data, as by adding them, all the standard deviation ellipses shrink dramatically comparing to the ellipses with doctor visit data only, standard deviation ellipses oval showed in Table 2.5. Infectious period data were also useful as the standard deviation ellipses shrinking by having them but with a much smaller magnitude.

Table 2.5: Oval Volume of Standard Deviation Oval for All Particle Filter Algorithms (×10000)

Oval Volume Comparison				
Algorithms	Doctor V.	Doctor V. & Infectious P.	Doctor V. & Contact T.	All
Bootstrap PF	8.3851	3.5917	4.2976	4.1926
PPF	0.5178	0.4204	0.1462	0.1177
APF	0.2608	0.6077	0.6574	1.6089
SSPPF1	1.6888	2.0318	0.7216	1.0057
SSPPF2	0.1848	0.1615	0.0972	0.0973

2.3.3 Particle Filter using Different Re-sampling Algorithm

In the above section, we also found that the performance of particle filter algorithms can be improved significantly by using improved re-sampling algorithms. In this section, we studied the importance of re-sampling algorithm by evaluating the performance of bootstrap particle filter, PPF and SSPPF2 using re-sampling algorithm described as Algorithm 2 in [1] (RA1) and Algorithm 4 (RA2) described in table 2.4 in this paper, respectively.

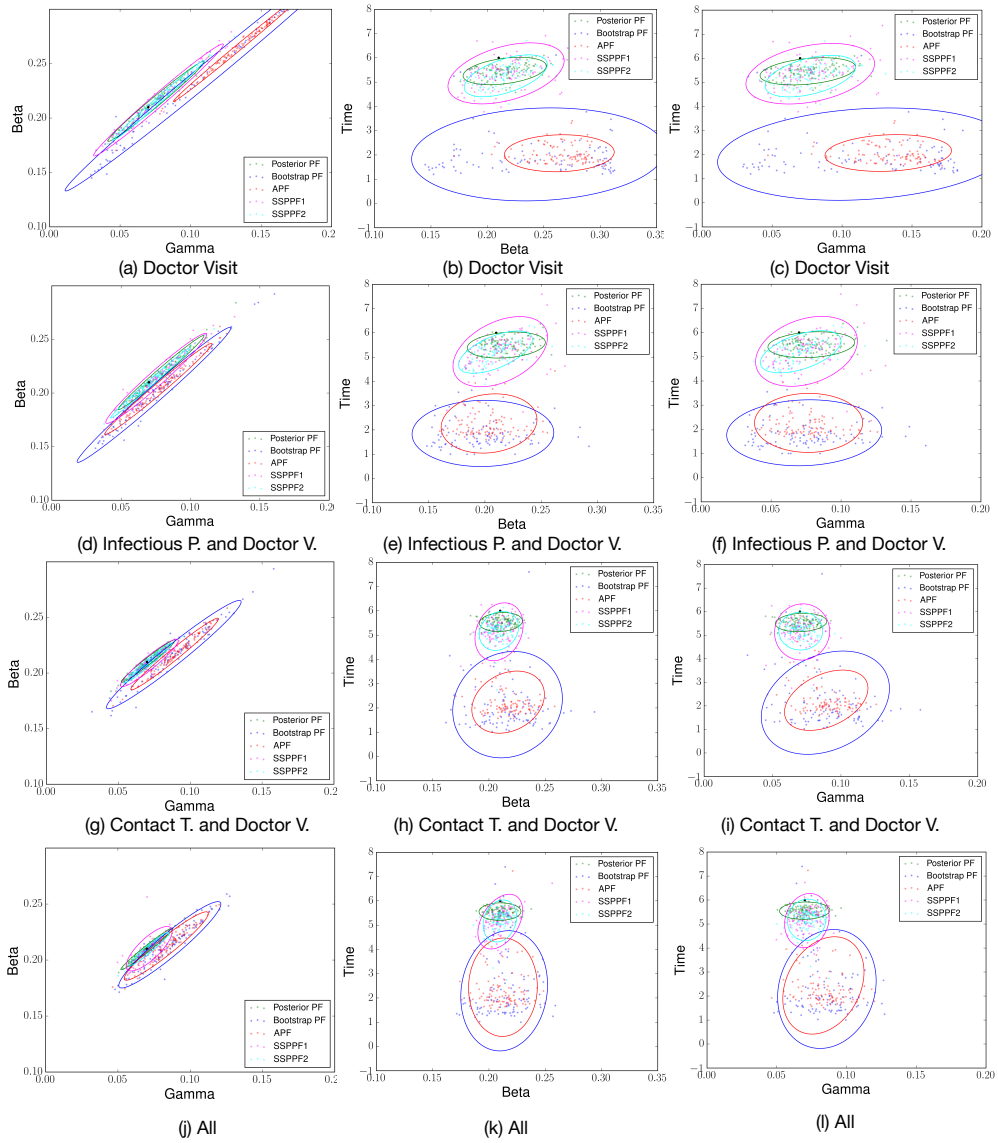


Figure 2.2: Fig (a) (d) (g) and (j) shows the standard deviation ellipses for 100 parameter estimations of β and γ for each particle filter algorithms with different data included. Fig (b) (e) (h) and (k) shows the standard deviation ellipses for 100 parameter estimations of β and t for each particle filter algorithms with different data included. Fig (c) (f) (i) and (l) shows the standard deviation ellipses for 100 parameter estimations of β and t for each particle filter algorithms with different data included.

RA2 performed significantly better in terms of stabilizing the particle filter algorithms. In table 2.6, Bootstrap PF, PPF, and SSPPF2 using RA2 provided a much smaller standard deviation oval comparing of using RA1 indicating a much smaller variations of estimations.

Table 2.6: Oval Volume of Standard Deviation Oval for Three Particle Filter Algorithms using RA1 and RA2 ($\times$10000)

Oval Volume Comparison		
Algorithms	RA1	RA2
Bootstrap PF	5.2449	1.6667
PPF	101.4522	0.1659
SSPPF2	121.9041	0.2442

2.3.4 Particle Filter using Different Number of Particles

Finally, we want to know whether the number of particles would influence the performance of the particle filter algorithms. From Table 2.7, bootstrap filter and APF performed comparable with either 400, 800, 1200 or 2400 particles. However, more particles could help stabilize the performance of PPF, SSPPF1, and SSPPF2 algorithms in finding the true β and γ values by providing smaller ovals with more particles.

2.4 Discussion

The compartmental SIR model we used is the most basic differential equation model for infectious disease progression. More complicated infectious

Table 2.7: Oval Volume of Standard Deviation Oval using Different Number of Particles ($\times$10000)

Oval Volume Comparison				
Algorithms	PF 300	PF 600	PF 1200	PF 2400
Bootstrap PF	7.6780	12.2794	4.1926	6.0496
PPF	1.2176	0.4965	0.1177	0.0638
APF	2.4744	2.2807	1.6089	2.0673
SSPPF1	5.6310	1.3992	1.0057	0.5614
SSPPF2	1.4495	0.2502	0.0973	0.1476

disease progression models can be found in [34]. The outbreak time parameter t is a parameter outside the differential equations of the compartmental SIR model, and the accuracy and consistency of the particle filter algorithms discussed in this paper may vary if we comparing more parameters inside infectious disease progression models.

Applying all historical data to adjust the weights of all particles does not hurt the computational efficiency. Our proposed algorithms applied all historical data in every iteration updating the weights of all particles, which intuitively thinking, should damage the computational efficiency. However, as the moving direction of all the particles were well restricted with the minimized effects of outliers, we found the algorithms actually run efficiently without wasting time having particles moving back and forward to look for the correct direction.

The number of iterations used in each particle filter run can influence the performance of the particle filters. The number of iterations for each

particle filter run in this paper was 40 iterations. For data set with no outliers, we believe more iterations will improve the accuracy and consistency of all the particle filters. However, for the data set with outliers, more iterations may lead to an inefficient performance of the particle filters as more outliers are applied and larger influence of them affecting the particles in finding the true parameters. Future work can be done to study an efficient way of determine the number of iterations applied in estimating the true parameters.

Chapter 3

Dengue lab diagnostic algorithms comparison

3.1 Introduction

Dengue is a global burden with an estimated 390 million dengue infections per year and 3.9 billion people in 128 countries at risk of infection and an estimated 500,000 people with severe dengue require hospitalization each year. [69]

Base on dengue virus (DENV) testing guidance provided by Centers for Disease Control and Prevention (CDC), dengue was suggested to be diagnosed by isolation of virus, by serological tests, or by molecular methods.

Historically dengue laboratory diagnosis base on single serum specimen with serum samples submitted during the first 5 days of symptoms (acute phase) tested using real-time reverse transcription polymerase chain reaction (RT-PCR) for molecular or non-structural protein 1 (NS1). RT-PCR is the primary tool to detect the DENV molecular in the early stage of the illness with a positive result as a proof of current infection and it also confirms the serotype of serum samples submitted. IgM antibody capture ELISA (MAC-ELISA) format is the most common diagnostic tool to capture human IgM antibodies for serum specimen submitted after 5 days (convalescent phase).

Recently, newly published tests sensitivity and specificity [30] motive CDC modifying test algorithm to be both molecular and IgM antibody tests (or NS1 and IgM antibody) in the acute phase (redefined acute phase to be 1-7 days) and IgM antibody tests in convalescent phase (redefined convalescent phase to be ≥ 7 days). In this chapter, we focus on the analysis of the different combinations performance of molecular (CDC RT-PCR) test, NS1 (InBios NS1 ELISA) and IgM test (CDC MAC ELISA) on both individual and population benefits.

The immune response varies depending on whether the individual has a primary (first dengue or other flavivirus infection) versus a secondary (had dengue or other flavivirus infection in past) dengue infection. [12] There is no statistically significant difference between primary and secondary dengue infection during the first 10 days of illness of detecting IgM antibody using CDC MAC ELISA. [30]

3.2 Materials and Methods

3.2.1 Assumptions

- Patients have their own sensitivities for RT-PCT, NS1 and IgM tests, respectively. The sensitivities for RT-PCT, NS1 and IgM tests are influenced by the virus RNA and human Antibody level. And sensitivity values are based on the Dengue diagnostic tests in single specimen diagnostic algorithm study in [30].

- The specificity are the same for all patients and not influenced by NAAT

and IgM levels.

- The test results of RT-PCT, NS1 and IgM will be independent of each other.

3.2.2 Prevalence Adjustment

Single test for specimen diagnostic algorithm Let T_i denote the test result for patient i, the diagnostic can be RT-PCT, NS1 or IgM test, with value equal to 1 and 0 represent positive test result and negative test result, respectively. Let $D = 1$ and $D = 0$ represent if the patient is dengue infected or not. The prevalence of the disease in the sample population is $p(D = 1) = \theta$. Let se_i represent the sensitivity of either RT-PCT, NS1 or IgM tests for patient i. And let sp_i represent the specificity of either RT-PCT, NS1 or IgM test for patient i.

The estimation of the prevalence, $\hat{\theta}$, is adjusted using the sensitivity and specificity of the each patients being tested to calculate the true sample prevalence θ according to the formula $\frac{\sum_{i=1}^{n} E(T_i(\theta))}{n} = \hat{\theta}$, with $E(T_i(\theta))$ represents the expectation of the test results of patient i. As $E(T_i(\theta)) = p(T_i = 1) \cdot 1 + p(T_i = 0) \cdot 0 = p(T_i = 1)$, with $p(T_i = 1)$ and $p(T_i = 0)$ represent positive test result and negative test result respectively, we target to find the expression of the probability for patient i to have positive test result.

The probability for patient to have dengue infected, $p(D = 1)$, is the true prevalence, denoted as θ. The probability for patient i to have a positive test result, $p(T_i = 1)$, for a simple test with sensitivity se_i and specificity sp_i

condition on the true prevalence θ is

$$\begin{aligned} p(T_i = 1) &= p(T_i = 1|D = 1)p(D = 1) + p(T_i = 1|D = 0)p(D = 0) \\ &= se_i \cdot \theta + (1 - sp_i) \cdot (1 - \theta) = \theta \cdot (se_i + sp_i - 1) + (1 - sp_i) \end{aligned} \tag{3.1}$$

The the adjusted prevalence estimation is

$$\theta = \frac{n \cdot \hat{\theta} - \sum_{i=1}^{n} B_i}{\sum_{i=1}^{n} A_i} \tag{3.2}$$

where

$$A_i = se_i + sp_i - 1$$

and

$$B_i = 1 - sp_i$$

with se_i and sp_i corresponding to the tests conducted.

Multiple tests for specimen diagnostic algorithm Current dengue lab testing algorithm adopt single specimen diagnostic algorithm with only one test performed for each specimen. According to the study [30], combining diagnostic tests by regarding at least one of the test positive as disease positive can improve the overall accuracy. In this paper, we analyze the cumulative test effects using the same algorithm in [30] by comparing the estimations and widths of the confidence intervals.

When 2 tests are used for diagnostic. The probability of having positive test results becomes

$$p(T_i = 1) = p(T_i^1 = 1 \cup T_i^2 = 1)$$

$$= (se_i^1 + se_i^2 - se_i^1 se_i^2 + sp_i^1 sp_i^2 - 1)\theta + 1 - sp_i^1 sp_i^2 \tag{3.3}$$

where T_i^1 and T_i^2 represent the test results from test 1 and 2, and test 1 and 2 can be RT-PCT, NS1 or IgM test. And se_i^1, se_i^2, sp_i^1 and sp_i^2 are the sensitivity and specificity associated with test 1 and 2, respectively.

When 3 tests are used for diagnostic. The probability of having positive test results becomes

$$p(T_i = 1) = p(T_i^1 = 1 \cup T_i^2 = 1 \cup T_i^3 = 1)$$

$$= (se_i^1 + se_i^2 + se_i^3$$

$$-se_i^1 se_i^2 - se_i^1 se_i^3 - se_i^2 se_i^3$$

$$+se_i^1 se_i^2 se_i^3 + sp_1 sp_2 sp_3 - 1)\theta \tag{3.4}$$

$$+1 - sp_1 sp_2 sp_3 \tag{3.5}$$

where T_i^1, T_i^2 and T_i^3 represent the test results from test 1, 2 and 3, and test 1,2,3 are RT-PCT, NS1 and IgM test. And se_i^1, se_i^2, se_i^3, sp_i^1, sp_i^2 and sp_i^3 are the sensitivity and specificity associated with test 1, 2, and 3, respectively.

3.2.3 Confidence Interval

We calculate the Sterne's confidence interval by inverting the test: $H_0 : \theta = \theta_c$ against $H_1 : \theta \neq \theta_c$, where θ and θ_c represent the unknown true

population prevalence and the adjusted prevalence estimations.

Base on H_0 and binomial distribution, the probability for k out of n patients to have disease is

$$\binom{n}{k}\theta_c^k(1-\theta_c)^{n-k}. \tag{3.6}$$

Given that there are k patients in the sample are with disease and assuming that each patient has their own sensitivity and specificity for the diagnostic tests, we can calculate the probability for m $(m = 0, 1, 2, ..., n)$ of them to be tested as disease positive base on Poisson binomial distribution [68] as:

$$\sum_{A\in F_m}\prod_{i\in A} p(T_i = 1)\prod_{i\in A^c}(1 - p(T_i = 1)), \tag{3.7}$$

where F_m is the set of all combinations of choosing m out of n samples and $p(T_i = 1)$ is calculated base on equations 3.93.33.4.

Thus, under $H_0 : \theta = \theta_c$, the probability of m patients are diagnostic as disease positive is:

$$P_{H_o}(\theta_c, m, n)$$
$$= \sum_{k=0}^{n}\left[\binom{n}{k}\theta_c^k(1-\theta_c)^{n-k}\Big(\sum_{A\in F_m}\prod_{i\in A} p(T_i = 1|k, n)\prod_{i\in A^c}(1 - p(T_i = 1|k, n))\Big)\right], \tag{3.8}$$

where

$$p(T = 1|k, n))$$
$$= se \cdot \frac{k}{n} + (1 - sp) \cdot (1 - \frac{k}{n}) = \frac{k}{n} \cdot (se + sp - 1) + (1 - sp) \tag{3.9}$$

The probability of obtaining a number of disease positive patients as probable as or less probable than a particular m number of disease positive patients (p-value) is

$$p_{value}(\theta_c, m, n) = \sum_{P_{H_o}(\theta_c,r,n) \leq P_{H_o}(\theta_c,m,n)} P_{H_o}(\theta_c, r, n), \tag{3.10}$$

where the summation is over all the values of r with $P_{H_o}(\theta_c, r, n) \leq P_{H_o}(\theta_c, m, n)$. For $(1 - \alpha)$ confidence interval, it is sufficient to consider the shortest interval of θ_c, that contains all the values of θ_c, for which

$$p_{value}(\theta_c, , n, m) \geq \alpha. \tag{3.11}$$

Observing m positive disease diagnostic, we can calculate the shortest confidence interval $\theta_1 \leq \theta_c \leq \theta_2$, which θ_1 is the smallest value satisfying equation 3.11 and θ_2 is the largest value satisfying equation 3.11. [59]

To save the computation time of Poisson Binomial distribution, we approximated the cumulative density function (cdf) with the cdf of refined standard normal distribution [65, 29]. The cdf of Poisson Binomial distribution is

$$F(x) = \sum_{m=0}^{x} \sum_{A \in F_m} \prod_{i \in A} p(T_i = 1) \prod_{i \in A^c} (1 - p(T_i = 1)) \quad for \quad x = 0, 1, 2, ...n. \tag{3.12}$$

The expectation, standard deviation and skewness are

$$\mu = \sum_{i=1}^{n} p(T_i = 1), \tag{3.13}$$

$$\sigma = (\sum_{i=1}^{n} p(T_i = 1)(1 - p(T_i = 1)))^{\frac{1}{2}}, \tag{3.14}$$

and

$$\gamma = \sigma^{-3} \sum_{i=1}^{n} p(T_i = 1)(1 - p(T_i = 1))(1 - 2p(T_i = 1)). \tag{3.15}$$

The cdf $F(x)$ is approximated as

$$F(x) = G(\frac{x + 0.5 - \mu}{\sigma}), \tag{3.16}$$

and $G(x) = \Phi(x) + \gamma(1 - x^2)\phi(x)/6$ with $\Phi(x)$ and $\phi(x)$ are the cdf and pdf of standard normal distribution with μ, σ and γ defined as equation 3.13, 3.14 and 3.15.

3.3 Results

In this section, we showed the performance of proposed test algorithms listed in Table 3.1. And we represent Molecular test using M, Dengue virus

antigen detection test (NS1) using NS1 and Serologic test using S. The first two algorithm $M\&S|S$ and $NS1\&S|S$ listed are the currently suggested algorithms by CDC. Algorithm $NS1|S$, $M\&NS1|S$, $M\&S$, $NS1\&S$, $M\&NS1\&S$ and $M\&NS1\&S|S$ are proposed algorithms. The proposed algorithms $NS1|S$, $M\&NS1|S$ and $NS1\&S$ were shown a better performance than currently suggested algorithms.

3.3.1 Data Simulation

We did 100 simulations. In each simulation run, we assumed the total population to be 10000 with the number of days after symptoms onset of each person randomly generated base on the distribution in [30] from 1 to 10 days. The sensitivity and specificity was randomly generated base on the normal assumption with the mean and standard error of the sensitivity and specificity from the study [30]. Tests results for each person are generated base on their probability of having a positive test result, $se_i^j = p(T_i^j = 1|D = 1)$ for infected patient and $(1 - sp_i^j) = p(T_i^j = 1|D = 0)$ for not infected patients, and the probability of having a negative test result, $(1 - se_i^j)p(T_i^j = 0|D = 1)$ for infected patient and $sp_i^j = p(T_i^j = 0|D = 0)$ for not infected patients. In each simulation run, 100 persons among 10000 was sampled randomly base on uniform distribution and tests were performed base on the number of days after symptom onset for each test algorithm listed in 3.1. For example, if two specimens were collected after 6 and 9 days of symptoms onset, respectively. If both tested using $M\&S|S$ test algorithm in table 3.1, molecular test and

serologic test will be performed for specimen with 6 days of symptom onset and serologic test will be performed for the 9 days specimen. For each sample in each simulation run, all test algorithms were performed for comparison. Results were analyzed base on 100 samples.

3.3.2 Population benefits

Prevalence Estimation Figure 3.1 shows the original estimation of prevalence using empirical method of calculating the prevalence ($p_{original} = \frac{n_{testpositives}}{n}$, where $n_{testpositives}$ and n are the number of specimens tested to be positive and total number of specimens tested in each sample.) and the prevalence after adjustment base on eq.3.2. The true prevalence is 0.1 and shown using red horizontal line. The empirical estimations tend to overestimate the prevalence due to the false positives happened when having low prevalence level, shown in dark blue box. Empirical estimations of algorithms $M\&NS1|S$, $M\&S$, $NS1\&S$, $M\&NS1\&S$ and $M\&NS1\&S|S$ having an higher estimations than algorithm $M\&S|S$ and $NS1\&S|S$. However, the adjustments can correct the empirical prevalence estimations of all algorithms to the right level shown using light blue box. With adjustments, all algorithms performs similar good in estimating prevalence.

Confidence Interval Width A narrower confidence interval is preferred for the test algorithm. Proposed algorithm $NS1|S$, $M\&NS1|S$, $NS1\&S$, $M\&NS1\&S$ and $M\&NS1\&S|S$ tend to have smaller confidence interval width

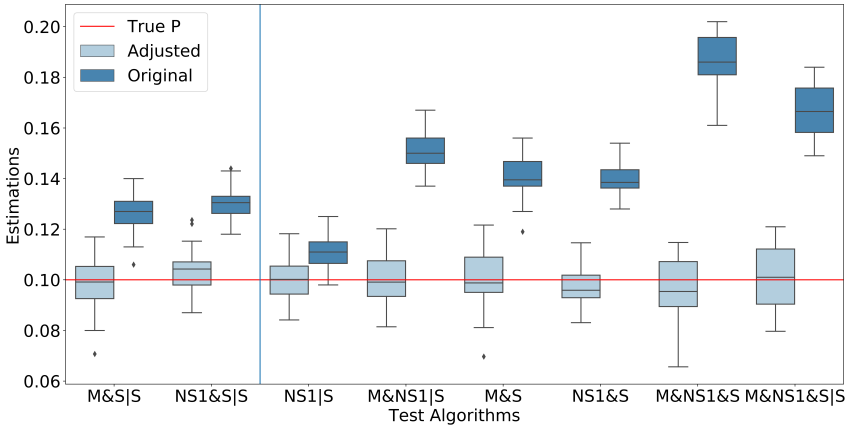


Figure 3.1: Figure shows the boxplots of original prevalence estimations (dark blue) and adjusted estimations (light blue). The true prevalence value ($p = 0.1$) shown using the horizontal red line.

than current suggested algorithm $M\&S|S$ and $NS1\&S|S$ base on boxplots of the confidence interval width in figure 3.2. Base on post-hoc test results 3.2, two currently suggested algorithms, $M\&S|S$ and $NS1\&S|S$, have similar confidence interval width with each other. Proposed algorithms $NS1|S$, $M\&NS1|S$, $NS1\&S$, and $M\&NS1\&S|S$ shows significant smaller confidence interval width comparing to $M\&S|S$ and $NS1\&S|S$. Proposed algorithm $M\&NS1\&S$ has not significant improved confidence interval width.

3.3.3 Individual Benefits

The current algorithm $NS1\&S|S$ performs better by giving higher sensitivity without sacrificing specificity comparing to $M\&S|S$ base on the post-hoc test results shown in Table 3.4 and 3.3. Algorithm $M\&NS1|S$, $NS1\&S$,

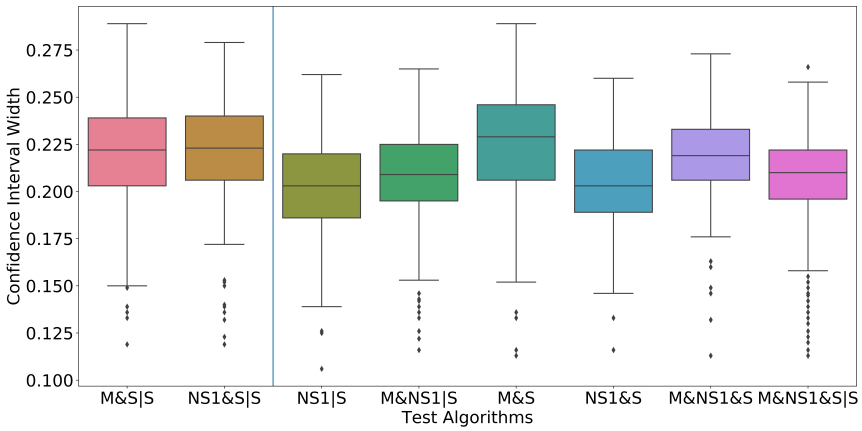


Figure 3.2: Figure shows the boxplots of the confidence interval widths of different algorithms.

and $M\&NS1\&S|S$ tend to increase the overall sensitivity without sacrificing overall specificity 3.3.

Algorithms $M\&NS1|S$, $NS1\&S$, $M\&NS1\&S$ and $M\&NS1\&S|S$, shown in Table 3.1, performs better than current suggested algorithms $M\&S|S$ with significant improved sensitivity 3.3. Algorithm $M\&NS1|S$ and $NS1\&S$ has similar sensitivity as current suggested algorithm $NS1\&S|S$, and $M\&NS1\&S$ and $M\&NS1\&S|S$ has significant better sensitivity than $NS1\&S|S$. Algorithm $NS1\&S$ has similar specificity as $M\&S|S$ but less specificity then $NS1\&S|S$. Algorithm $M\&NS1|S$, $M\&NS1\&S$ and $M\&NS1\&S|S$ has less specificity comparing to $M\&S|S$ and $NS1\&S|S$. However, the decrease of specificity are much smaller than the increase of sensitivity of algorithm $M\&NS1|S$, $NS1\&S$, $M\&NS1\&S$ and $M\&NS1\&S|S$, shown in table 3.5.

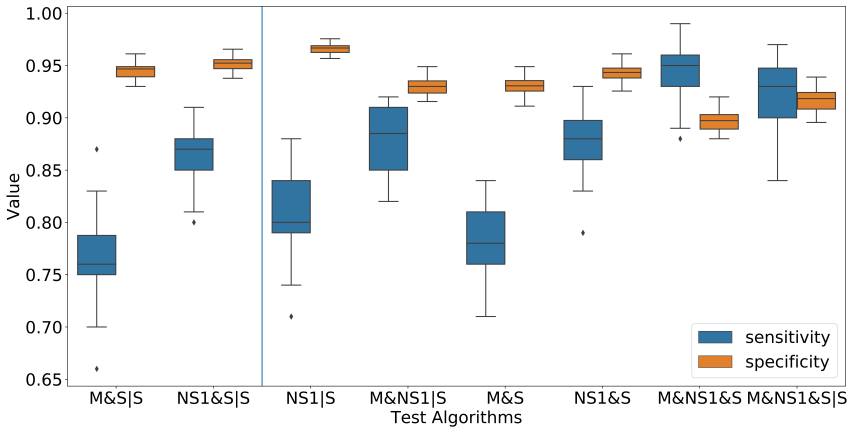


Figure 3.3: Figure shows the sensitivity and specificity of each algorithm for individual level benefits.

3.4 Discussion

The proposed and current suggested algorithms tend to over estimate the prevalence with more false positive. However, with the adjustment base on eq.3.2, all proposed algorithms and current suggested algorithms can be adjusted to the correct prevalence level.

Algorithm $NS1|S$ can be an alternative of currently suggested algorithm $NS1\&S|S$, base on table 3.5. $NS1\&S|S$ involves doing both NS1 and Serologic tests for patients with less than 7 days symptoms onset, while $NS1|S$ only require a single NS1 test for patients with less than 7 days symptoms onset. For patients with more than 7 days after symptom onset, $NS1|S$ and $NS1\&S|S$ both only use Serologic test. $NS1|S$ test is simpler than $NS1\&S|S$ test with only one test conducted for acute phase patient. The overall sensi-

tivity of $NS1|S$ is about 6% worse than $NS1\&S|S$ but with a similar overall specificity as $NS1\&S|S$. The confidence interval of $NS1|S$ algorithm is about 91% width of the confidence interval of $NS1\&S|S$.

Algorithm $M\&NS1|S$ and $NS1\&S$ have increased overall sensitivity and narrowed confidence interval with slightly decreased specificity comparing to currently suggested $M\&S|S$ and $NS1|S$. Algorithm $M\&NS1\&S$ and $M\&NS1\&S|S$ have further increased overall sensitivity and similar confidence interval comparing to $M\&NS1|S$ and $NS1\&S$ similar but around $3\% - 5\%$ decrease in specificity. Algorithm $M\&NS1|S$ has same number of tests as suggested algorithm with the replacement Serologic test with NS1. Algorithm $NS1\&S$ has one more test conducted in convalescent phase. As we only studied $1 - 10$ days after symptom onset, the sensitivity and specificity of NS1 test might decrease significantly after 10 days. Algorithm $M\&NS1\&S$ and $M\&NS1\&S|S$ requiring one more tests to be conducted in acute phase, and $M\&NS1\&S$ require two more tests in convalescent phase which will increase of cost dramatically. Considering both the performance and complexity of algorithm, $NS1|S$ and $M\&NS1|S$ are suggested as $NS1|S$ is simpler but with similar performance and $M\&NS1|S$ has same complexity but with better overall performance.

Table 3.1: Dengue testing Algorithms

Dengue Testing Algorithms		
Notation	Algorithms	Explanation
$M\&S\|S$	Molecular and Serologic or Serologic	Molecular test and Serologic test for specimens collected < 7 days after symptom onset, or Serologic test for specimens collected ≥ 7 days after symptom onset
$NS1\&S\|S$	NS1 and Serologic or Serologic	NS1 test and Serologic test for specimens collected < 7 days afater symptom onset, or Serologic test for specimens collected ≥ 7 days afater symptom onset
$NS1\|S$	NS1 or Serologic	NS1 test for specimens collected < 7 days or Serologic test for specimens collected ≥ 7 days afater symptom onset
$M\&NS1\|S$	Molecular and NS1 or Serologic	Molecular test and NS1 test for specimens collected < 7 days afater symptom onset, or Serologic test for specimens collected ≥ 7 days afater symptom onset
$M\&S$	Molecular and Serologic	Molecular test and Serologic test for specimens collected ≥ 1 days
$NS1\&S$	NS1 and Serologic	NS1 test and Serologic test for specimens collected ≥ 7 days
$M\&NS1\&S$	Molecular and NS1 and Serologic	Molecular and NS1 test and Serologic test for specimens collected ≥ 1 days afater symptom onset
$M\&NS1\&S\|S$	Molecular and NS1 and Serologic or Serologic	Molecular and NS1 test and Serologic test for specimens collected < 7 days afater symptom onset, or Serologic test for specimens collected ≥ 7 days afater symptom onset

Table 3.2: Post-hoc test results for confidence interval widths

Post-hoc test results for confidence interval widths		
contrast (proposed − current)	estimate	p-value
$NS1\&S\|S$(NS1 and Serologic or Serologic) − $M\&S\|S$(Molecular and Serologic or Serologic)	0.004	0.8262
$NS1\|S$(NS1 or Serologic)− $M\&S\|S$	-0.016	$< .0001$
$NS1\|S$(NS1 or Serologic)− $NS1\&S\|S$	-0.019	$< .0001$
$M\&NS1\|S$(Molecular and NS1 or Serologic)− $M\&S\|S$	-0.010	0.0005
$M\&NS1\|S$(Molecular and NS1 or Serologic)− $NS1\&S\|S$	-0.014	$< .0001$
$NS1\&S$(NS1 and Serologic)− $M\&S\|S$	-0.018	$< .0001$
$NS1\&S$(NS1 and Serologic)− $NS1\&S\|S$	-0.016	$< .0001$
$M\&NS1\&S$(Molecular and NS1 and Serologic)− $M\&S\|S$	-0.002	0.9968
$M\&NS1\&S$(Molecular and NS1 and Serologic)− $NS1\&S\|S$	-0.006	0.304
$M\&NS1\&S\|S$(Molecular and NS1 and Serologic or Serologic)− $M\&S\|S$	-0.01 -0.01	0.0019 0.0019
$M\&NS1\&S\|S$(Molecular and NS1 and Serologic or Serologic)− $NS1\&S\|S$	-0.013 -0.013	$< .0001$ $< .0001$

Table 3.3: Post-hoc test results of sensitivity

Post-hoc test results of sensitivity		
contrast (proposed − current)	estimate	p-value
$NS1\&S\|S$(NS1 and Serologic or Serologic) − $M\&S\|S$(Molecular and Serologic or Serologic)	0.0986	$< .0001$
$M\&NS1\|S$(Molecular and NS1 or Serologic)− $M\&S\|S$	0.1110	$< .0001$
$M\&NS1\|S$(Molecular and NS1 or Serologic)− $NS1\&S\|S$	0.012	0.9248
$NS1\&S$(NS1 and Serologic)− $M\&S\|S$	0.114	$< .0001$
$NS1\&S$(NS1 and Serologic)− $NS1\&S\|S$	0.015	0.7815
$M\&NS1\&S$(Molecular and NS1 and Serologic)− $M\&S\|S$	0.178	$< .0001$
$M\&NS1\&S$(Molecular and NS1 and Serologic)− $NS1\&S\|S$	0.079	$< .0001$
$M\&NS1\&S\|S$(Molecular and NS1	0.152	$< .0001$
and Serologic or Serologic)− $M\&S\|S$	0.152	$< .0001$
$M\&NS1\&S\|S$(Molecular and NS1	0.0533	$< .0001$
and Serologic or Serologic)− $NS1\&S\|S$	0.0533	$< .0001$

Table 3.4: Post-hoc test results for specificity

Post-hoc test results for specificity		
contrast (proposed − current)	estimate	p-value
$NS1\&S\|S$(NS1 and Serologic or Serologic) −A5(Molecular and Serologic or Serologic)	0.0067	0.0373
$M\&NS1\|S$(Molecular and NS1 or Serologic)− $M\&S\|S$	−0.0147	$< .0001$
$M\&NS1\|S$(Molecular and NS1 or Serologic)− $NS1\&S\|S$	−0.021	$< .0001$
$NS1\&S$(NS1 and Serologic)− $M\&S\|S$	−0.0016	0.9972
$NS1\&S$(NS1 and Serologic)− $NS1\&S\|S$	−0.0083	0.0025
$M\&NS1\&S$(Molecular and NS1 and Serologic)− $M\&S\|S$	−0.047	$< .0001$
$M\&NS1\&S$(Molecular and NS1 and Serologic)− $NS1\&S\|S$	−0.0054	$< .0001$
$M\&NS1\&S\|S$(Molecular and NS1	−0.027	$< .0001$
and Serologic or Serologic)− $M\&S\|S$	−0.027	$< .0001$
$M\&NS1\&S\|S$(Molecular and NS1	−0.034	$< .0001$
and Serologic or Serologic)− $NS1\&S\|S$	−0.034	$< .0001$

Table 3.5: Algorithm performance comparison

Algorithm performance comparison						
Algorithm	Sensitivity	change	Specificity	change	CI width	%change
$M\&S \mid S$(Molecular and Serologic or Serologic)	0.766		0.944		0.216	
$NS1\&S \mid S$(NS1 and Serologic or Serologic)	0.865	9.9%	0.951	0.7%	0.220	1.85%
$NS1 \mid S$(NS1 or Serologic)	0.805	3.9%	0.966	2.2%	0.201	−6.9%
$M\&NS1 \mid S$ (Molecular and NS1 or Serologic)	0.877	11.1%	0.93	−1.4%	0.206	−4.6%
$M\&S$(Molecular and Serologic)	0.782	1.6%	0.931	−1.3%	0.222	2.6%
$NS1\&S$(NS1 and Serologic)	0.880	11.4%	0.943	−0.1%	0.204	−5.56%
$M\&NS1\&S$ (Molecular and NS1 and Serologic)	0.944	17.8%	0.897	−4.7%	0.214	−0.9%
$M\&NS1\&S \mid S$ (Molecular and NS1 and Serologic or Serologic)	0.918	15.2%	0.917	−2.7%	0.207	−4.17%

Appendices

Appendix A

A.1 Analytic Expression of Uncertainty

Consider a region with geographic divisions given by $X = \{x_1, x_2, ..., x_n\}$. Suppose some species lives in the region, and the fraction of the species that lives in division i is p_i. A basic goal in SDM is to reconstruct the geographic distribution $P = \{p_1, p_2, ..., p_n\}$. To do this, we have some species occurrence data $O = \{o_1, o_2, ..., o_n\}$, where each o_i specifies the number of times the species has occurred in division i. The occurrence data can be viewed as a sample from the distribution P. In addition we are given k layers of environmental data for the region described by features $f_j(X)$ for $j = 1, \ldots, k$. For example, one such function could be the average elevation in each geographic division.

The mathematical formulation of the maximum entropy problem is

$$\max_{p_i} \; -\sum_{i=1}^{n} p_i \log p_i \tag{A.1a}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} p_i f_j(x_i) = \hat{E}(f_j(X)) \quad j = 1, \ldots, k \tag{A.1b}$$

$$\sum_{i=1}^{n} p_i = 1 \tag{A.1c}$$

$$p_i \geq 0 \qquad i = 1, \ldots, n \tag{A.1d}$$

The counts $O = \{o_1, o_2, ...o_n\}$ follow a multinomial distribution, whose

true parameter $P = \{p_1, p_2, ..., p_n\}$ is unknown. Let $m = \sum_{i=1}^{n} o_i$ be the total number of observations. The maximum likelihood estimator of $P = \{p_1, p_2, ..., p_n\}$, is

$$\tilde{P} = \begin{pmatrix} \tilde{p}_1 \\ \tilde{p}_2 \\ \vdots \\ \tilde{p}_n \end{pmatrix} = \frac{1}{m} \begin{pmatrix} o_1 \\ o_2 \\ \vdots \\ o_n \end{pmatrix} = \frac{O}{m}$$

and it follows a normal distribution [67],

$$\tilde{P} \sim \text{Normal}(P, \frac{\mathbf{\Sigma}}{m}),$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_n \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_n \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_n & -p_2p_n & \cdots & p_n(1-p_n) \end{pmatrix}.$$

From the analysis in Kapur et al. [32], one can show that the maximum likelihood estimates of P from maximum entropy model (A.1) are achieved when

$$\hat{E}(f_j(X)) = \frac{\sum_{i=1}^{n} f_j(x_i) o_i}{\sum_{i=1}^{n} o_i} \tag{A.2}$$

For brevity, let $a_j = \hat{E}(f_j(X))$. Based on equation (A.2), the vector of a_j can be expressed as

$$A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} = \frac{1}{m} \begin{pmatrix} f_1(x_1) & f_1(x_2) & \cdots & f_1(x_n) \\ f_2(x_1) & f_2(x_2) & \cdots & f_2(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ f_k(x_1) & f_k(x_2) & \cdots & f_k(x_n) \end{pmatrix} \begin{pmatrix} o_1 \\ o_2 \\ \vdots \\ o_n \end{pmatrix} \tag{A.3}$$

Let

$$\mathbf{F} = \begin{pmatrix} f_1(x_1) & f_1(x_2) & \cdots & f_1(x_n) \\ f_2(x_1) & f_2(x_2) & \cdots & f_2(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ f_k(x_1) & f_k(x_2) & \cdots & f_k(x_n) \end{pmatrix},$$

then

$$A = \frac{1}{m} \cdot \mathbf{F} \cdot O = \mathbf{F} \cdot \tilde{P}$$

Because $\mathbf{F}$ is constant, A is an affine transformation of $\tilde{P}$. Using the distribution of $\tilde{P}$, we can write the distribution of A [5]

$$A \sim \text{Normal}(\mathbf{F} \cdot P, \frac{\mathbf{F} \cdot \mathbf{\Sigma} \cdot \mathbf{F}^T}{m}). \tag{A.4}$$

Let $g(A)$ denote the maximum entropy optimization, model (A.1), as a function from $\mathcal{R}^k$ to $\mathcal{R}^n$. In other words, the function takes as input the vector A, specifying right hand sides of the equality constraints $\hat{E}(f_j(X))$, and outputs a probability estimate across the geographic region P. We would like to understand the uncertainty in the output $g(A)$ as a function of the uncertainty of the input A. This can be done following steps similar to those in the delta method [2, p.75].

To understand the uncertainty in the output $g(A)$, we begin by writing a first order Taylor expansion of g around $E(A)$

$$\begin{aligned} g(A) &\approx g(E(A)) + \nabla g(E(A)) \cdot [A - E(A)] \\ &\approx g(\mathbf{F} \cdot P) + \nabla g(\mathbf{F} \cdot P) \cdot [A - E(A)], \end{aligned} \tag{A.5}$$

where $\nabla g(\cdot)$ is an $n \times k$ matrix of partial derivatives, with entry (i, j) specified by $\frac{\partial p_i}{\partial a_j}$. If we can compute an expression for these partial derivatives, then everything on the right hand side above is constant, except $[A - E(A)]$ whose distribution we know because we know the distribution of A. $g(A)$ is an affine transformation of $[A - E(A)]$, and can be approximated as

$$g(A) \sim \text{Normal}(g(\mathbf{F} \cdot P), \nabla g \cdot \frac{\mathbf{F} \cdot \mathbf{\Sigma} \cdot \mathbf{F}^T}{m} \cdot (\nabla g)^T). \tag{A.6}$$

To complete the analysis of the output uncertainty, we continue by deriving an expression for $\frac{\partial p_i}{\partial a_j}$. We introduce some additional notation, following Kapur et al. [32]. Let λ be the Lagrange multiplier for constraint (A.1c), and μ_j be the multiplier for constraint (A.1b) for $j = 1, \ldots, k$. It can be shown [32] that the optimal p_i have the expression

$$p_i = e^{-\sum_{j=1}^{k} \mu_j f_j(x_i) - \lambda - 1} \qquad \forall i = 1, 2, ...n. \tag{A.7}$$

Using the constraint (A.1c), we have

$$\sum_{t=1}^{n} e^{-\lambda - 1 - \sum_{j=1}^{k} f_j(x_t)\mu_j} = 1$$

$$e^{\lambda+1} = \sum_{t=1}^{n} e^{-\sum_{j=1}^{k} f_j(x_t)\mu_j}. \tag{A.8}$$

We can now substitute the expression for $e^{\lambda+1}$ back into (A.7) to derive

$$p_i = \frac{e^{-\sum_{j=1}^{k} f_j(x_i)\mu_j}}{\sum_{t=1}^{n} e^{-\sum_{j=1}^{k} f_j(x_t)\mu_j}} \tag{A.9}$$

We now have an expression of the p_i in terms of the dual multipliers μ_j. But, we would like to compute $\frac{\partial p_i}{\partial a_j}$, which we can do by first computing partial derivatives with respect to the μ_j and using the expression

$$\frac{\partial p_i}{\partial a_j} = \sum_{r=1}^{k} \frac{\partial p_i}{\partial \mu_r} \frac{\partial \mu_r}{\partial a_j}.$$

What remains to be computed is $\frac{\partial p_i}{\partial \mu_r}$ and $\frac{\partial \mu_r}{\partial a_j}$. From (A.9), we have

$$\begin{aligned}
\frac{\partial p_i(\mu_1, \mu_2 ... \mu_m)}{\partial \mu_r} &= -f_r(x_i) e^{-\sum_{j=1}^{k} f_j(x_i)\mu_j} (\sum_{t=1}^{n} e^{-\sum_{j=1}^{k} f_j(x_t)\mu_j})^{-1} \\
&\quad - (\sum_{t=1}^{n} e^{-\sum_{j=1}^{k} f_j(x_t)\mu_j})^{-2} (\sum_{t=1}^{n} (-1) f_r(x_t) e^{-\sum_{j=1}^{k} f_j(x_t)\mu_j}) e^{-\sum_{j=1}^{k} f_j(x_i)\mu_j} \\
&= -f_r(x_i) p_i - p_i (\sum_{t=1}^{n} (-1) f_r(x_t) p_t) \\
&= -f_r(x_i) p_i + p_i a_r \\
&= p_i (a_r - f_r(x_i)),
\end{aligned}$$

where we derived the first equality from the chain rule, the second equality by substituting using expression (A.9), and the third equality by using the fact that $a_r = \sum_{t=1}^{n} f_r(x_t) p_t$ because p_i's are feasible and optimal in the maximum entropy optimization.

Then, we want to get the value of $\frac{\partial \mu_r}{\partial a_j}$. It is hard to get the expression of μ_r in terms of a_j, however, we can derive $\frac{\partial a_j}{\partial \mu_r}$ and use the Inverse Function Theorem [58] to calculate $\frac{\partial \mu_r}{\partial a_j}$. To get the expression of a_j in terms of μ_r, we substitute the expression of the optimal p_i (A.7) into constraint (A.1b), and then plug-in the expression of λ in terms of μ_r (A.8). Then, we express a_j as

$$a_j = \frac{\sum_{t=1}^{n} f_j(x_t) e^{-\sum_{j=1}^{k} f_j(x_t)\mu_j}}{\sum_{t=1}^{n} e^{-\sum_{j=1}^{k} f_j(x_t)\mu_j}}. \tag{A.10}$$

From A.10, we have

$$\frac{\partial a_j}{\partial \mu_r} \tag{A.11}$$

$$= \sum_{i=1}^{n}(-f_r(x_i)) f_j(x_i) e^{-\sum_{j=1}^{m} f_j(x_i)\mu_j} (\sum_{i=1}^{n} e^{-\sum_{j=1}^{k} f_j(x_i)\mu_j})^{-1} \tag{A.12}$$

$$- (\sum_{i=1}^{n} e^{-\sum_{j=1}^{m} f_j(x_i)\mu_j})^{-2} (\sum_{i=1}^{n} f_j(x_i) e^{-\sum_{j=1}^{k} f_j(x_i)\mu_j}) (\sum_{i=1}^{n} -f_r(x_i) e^{-\sum_{j=1}^{k} f_j(x_i)\mu_j}) \tag{A.13}$$

$$= -\sum_{i=1}^{n} f_r(x_i) f_j(x_i) p_i + (\sum_{i=1}^{n} f_j(x_i) p_i)(\sum_{i=1}^{n} f_r(x_i) p_i) \tag{A.14}$$

$$= -cov_P(f_r, f_j), \tag{A.15}$$

where we derived the first equality from the chain rule, the second equality by substituting using expression (A.9). The final equality comes from the definition of covariance, where we take the covariance of feature f_r and f_j with respect to the maximum entropy model results p_i.

If the determinant of the covariance is non-zero, following the Inverse Function Theorem [58], the inverse is differentiable. We denote the covariance matrix of features with respect to the maximum entropy model results as $\mathbf{\Psi}$, where $\mathbf{\Psi}_{rj} = cov_P(f_r, f_j)$. We denote the inverse covariance matrix as $\mathbf{\Psi}^{-1}$ and refer to its (r, j)th entry as $(\mathbf{\Psi}^{-1})_{rj}$. By the inverse function theorem, $\frac{\partial \mu_r}{\partial a_j}$

is equal to $(\mathbf{\Psi}^{-1})_{rj}$. Finally, we can express the $\frac{\partial p_i}{\partial a_j}$ as

$$\frac{\partial p_i}{\partial a_j} = \sum_{r=1}^{k} p_i(a_r - f_r(x_i))(-\mathbf{\Psi}^{-1})_{rj}. \tag{A.16}$$

A.2 Increasing programming speed

In the calculation of the relative probability for Aedes aegypti for a $1km^2$ square grid, we have 933,680 grid cells in total. The problem of computing the covariance of the output mainly comes from $\mathbf{\Sigma}$ where

$$\mathbf{\Sigma} = \begin{pmatrix} \hat{p}_1(1-\hat{p}_1) & -\hat{p}_1\hat{p}_2 & \cdots & -\hat{p}_1\hat{p}_n \\ -\hat{p}_1\hat{p}_2 & \hat{p}_2(1-\hat{p}_2) & \cdots & -\hat{p}_2\hat{p}_n \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{p}_1\hat{p}_n & -\hat{p}_2\hat{p}_n & \cdots & \hat{p}_n(1-\hat{p}_n) \end{pmatrix}.$$

The matrix is of size 933680×933680, which can cause out of memory errors.

To figure out a way of speeding the calculation, we first split $\mathbf{\Sigma}$ into two parts, where

$$\begin{aligned} \mathbf{\Sigma} &= \begin{pmatrix} \hat{p}_1(1-\hat{p}_1) & -\hat{p}_1\hat{p}_2 & \cdots & -\hat{p}_1\hat{p}_n \\ -\hat{p}_1\hat{p}_2 & \hat{p}_2(1-\hat{p}_2) & \cdots & -\hat{p}_2\hat{p}_n \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{p}_1\hat{p}_n & -\hat{p}_2\hat{p}_n & \cdots & \hat{p}_n(1-\hat{p}_n) \end{pmatrix} \\ &= \begin{pmatrix} -\hat{p}_1^2 & -\hat{p}_1\hat{p}_2 & \cdots & -\hat{p}_1\hat{p}_n \\ -\hat{p}_1\hat{p}_2 & -\hat{p}_2^2 & \cdots & -\hat{p}_2\hat{p}_n \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{p}_1\hat{p}_n & -\hat{p}_2\hat{p}_n & \cdots & -\hat{p}_n^2 \end{pmatrix} + \begin{pmatrix} \hat{p}_1 & 0 & \cdots & 0 \\ 0 & \hat{p}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{p}_n \end{pmatrix}. \end{aligned}$$

Define

$$\mathbf{\Sigma}_1 = \begin{pmatrix} -\hat{p_1}^2 & -\hat{p_1}\hat{p_2} & \cdots & -\hat{p_1}\hat{p_n} \\ -\hat{p_1}\hat{p_2} & -\hat{p_2}^2 & \cdots & -\hat{p_2}\hat{p_n} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{p_1}\hat{p_n} & -\hat{p_2}\hat{p_n} & \cdots & -\hat{p_n}^2 \end{pmatrix},$$

and

$$\mathbf{\Sigma}_2 = \begin{pmatrix} \hat{p_1} & 0 & \cdots & 0 \\ 0 & \hat{p_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{p_n} \end{pmatrix}.$$

The covariance of the output P can be estimated as

$$\nabla g \cdot \frac{\mathbf{F} \cdot \mathbf{\Sigma} \cdot \mathbf{F}^T}{m} \cdot (\nabla g)^T = \nabla g \cdot \frac{\mathbf{F} \cdot (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2) \cdot \mathbf{F}^T}{m} \cdot (\nabla g)^T.$$

First, to calculate $\nabla g \cdot \frac{\mathbf{F} \cdot \mathbf{\Sigma}_1 \cdot \mathbf{F}^T}{m} \cdot (\nabla g)^T$, we define

$$P_1 = \begin{pmatrix} \hat{p_1} & 0 & \cdots & 0 \\ \hat{p_2} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{p_n} & 0 & \cdots & 0 \end{pmatrix},$$

then we have

$$\nabla g \cdot \frac{\mathbf{F} \cdot \mathbf{\Sigma}_1 \cdot \mathbf{F}^T}{m} \cdot (\nabla g)^T = \nabla g \cdot \frac{\mathbf{F} \cdot P_1 \cdot P_1^T \cdot \mathbf{F}^T}{m} \cdot (\nabla g)^T.$$

The i^{th} diagonal element of $\nabla g \cdot \mathbf{F} \cdot \Sigma_1 \cdot \mathbf{F}^T \cdot (\nabla g)^T$, denoted as d_i^1, can be calculated as

$$\begin{aligned} d_i^1 &= \left(\frac{\partial p_i}{\partial a_1}(f_1(x_1)\hat{p_1} + \cdots + f_1(x_n)\hat{p_n}) + \cdots + \frac{\partial p_i}{\partial a_k}(f_k(x_1)\hat{p_1} + \cdots + f_k(x_n)\hat{p_n}) \right)^2 \\ &= \left(\sum_{\ell=1}^{k} \frac{\partial p_i}{\partial a_\ell} (\sum_{m=1}^{n} f_\ell(x_m)\hat{p_m}) \right)^2 \end{aligned} \tag{A.17}$$

To calculate $\nabla g \cdot \frac{\mathbf{F}\cdot\mathbf{\Sigma}_2\cdot\mathbf{F}^T}{m} \cdot (\nabla g)^T$, we define

$$P_2 = \begin{pmatrix} \sqrt{\hat{p}_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\hat{p}_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\hat{p}_n} \end{pmatrix},$$

then we have

$$\nabla g \cdot \frac{\mathbf{F}\cdot\mathbf{\Sigma}_2\cdot\mathbf{F}^T}{m} \cdot (\nabla g)^T = \nabla g \cdot \frac{\mathbf{F}\cdot P_2 \cdot P_2^T \cdot \mathbf{F}^T}{m} \cdot (\nabla g)^T.$$

The i^{th} diagonal element of $\nabla g \cdot \mathbf{F} \cdot \Sigma_2 \cdot \mathbf{F}^T \cdot (\nabla g)^T$, denoted as d_i^1, is calculated by

$$\begin{aligned} d_i^2 = &(\frac{\partial p_i}{\partial a_1} f_1(x_1)\sqrt{\hat{p_1}} + \cdots + \frac{\partial p_i}{\partial a_k} f_k(x_1)\sqrt{\hat{p_1}})^2 + \cdots + \\ &(\frac{\partial p_i}{\partial a_1} f_1(x_n)\sqrt{\hat{p_n}} + \cdots + \frac{\partial p_i}{\partial a_k} f_k(x_n)\sqrt{\hat{p_n}})^2 \\ = &\sum_{m=1}^{n} \left(\sum_{\ell=1}^{k} \frac{\partial p_i}{\partial a_\ell} f_\ell(x_m)\sqrt{\hat{p_m}} \right)^2 \end{aligned} \tag{A.18}$$

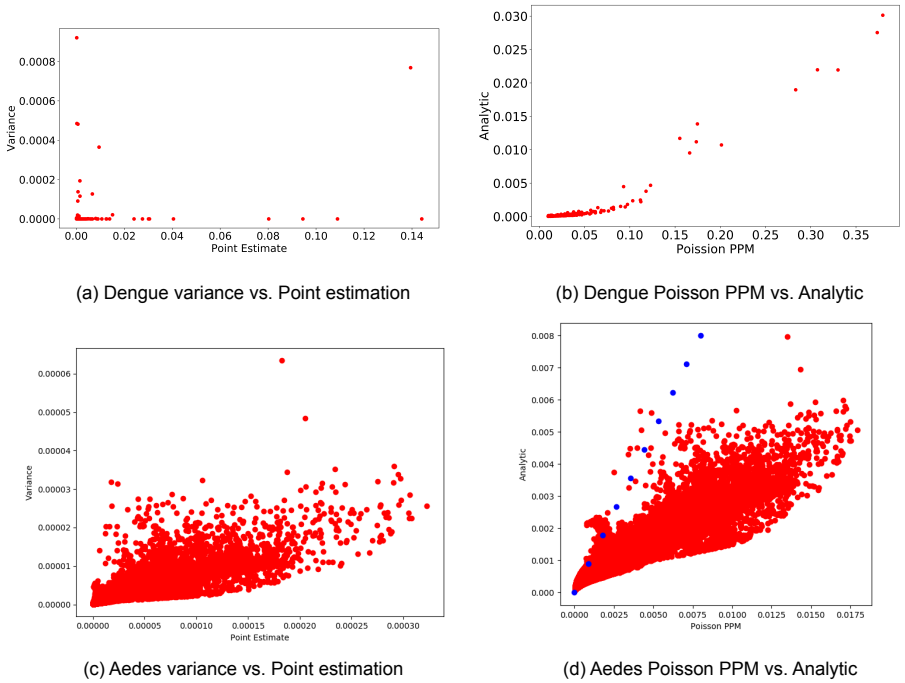
After we have calculated both d_i^1 and d_i^2 based on A.17 and A.18, we can calculate the variance of the p_i as $\frac{(d_i^1+d_i^2)}{m}$.

A.3 Comparison between Analytic method and Poisson PPM

Poisson PPM was proved to be equivalent to maximum entropy model with hidden assumptions of independence data. We showed the plots of variance vs. point estimations and standard deviation comparison between Poisson PPM and analytic method for both Dengue importation probability and

Aedes Aegypti suitability probability. Fig A.1a shows the relationship between Dengue importation probability point estimates and estimated variance using analytic method, which indicating the possible improper use of Poisson PPM approach. Fig A.1b shows a standard deviation comparison between the analytic and Poisson PPM method for Dengue importation probability. The regression line between two results is $s_a = 0.054 s_p$ with $R^2 = 0.805$, where s_a and s_p stand for the standard deviation estimates from the analytic and Poisson PPM methods, respectively. Poisson PPM gives a much larger standard deviation comparing to analytic and bootstrap method indicating the possible violate of the case location independence assumption.

Fig A.1c shows the relationship between Aedes Aegypti suitability point estimates and estimated variance using analytic method. There is more linear relationship comparing to Dengue importation cases. Fig A.1d shows the standard deviation comparison between analytic method and Poisson PPM approach for Aedes Aegypti suitability. Each red dot represent the standard deviation estimates for each grid using Poisson PPM and analytic method respectively. The blue dot shows the diagonal line when two methods aligned well. We have the relationship $s_a = 0.0917 s_p$ with $R^2 = 0.812$, where s_a and s_p stand for the standard deviation estimates from the analytic and Poisson PPM methods, respectively. Similar as Dengue case, Poisson PPM doesn't seem to functional well with a much larger standard deviation estimation comparing to analytic and bootstrap method.



(a) Dengue variance vs. Point estimation

(b) Dengue Poisson PPM vs. Analytic

(c) Aedes variance vs. Point estimation

(d) Aedes Poisson PPM vs. Analytic

Figure A.1: **Analytic and Poisson PPM Comparison** (a) Figure plots the relationship between point estimates of Dengue importation probability vs. variance calculated through analytic method. Non-linear relationship indicates the improper use of Poisson PPM for Dengue importation cases. (b) Figure plots the standard deviations of Poisson PPM vs. analytic for Dengue importation case study and indicates that Poisson PPM provides much larger standard deviation for Dengue imports application. (c) Figure plots the relationship between point estimates of Aedes Aegypti existence probability vs. variance calculated through analytic method. (d) Figure shows the standard deviation comparison between analytic method and Poisson PPM of Aedes Aegypti existence probability.

Bibliography

[1] M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on signal processing*, 50(2):174–188, 2002.

[2] Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer Science & Business Media, 2007.

[3] MP Austin. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological modelling*, 157(2):101–118, 2002.

[4] Carolyn Elizabeth Barney. Dengue risk factor distribution in harris county, texas. Master's thesis, M.Sc. Thesis, The University of Texas School of Public Health, 2008. Accessed: 2016-04-28.

[5] Marcel Berger. *Geometry i*. Springer Science & Business Media, 2009.

[6] Frida Cano, Jaycob Gorski, and Jessica Eastridge. Mosquito surveillance in the brazos county (diptera: Culicidae). *Instars: A Journal of Undergraduate Research*, 1(1), 2015.

[7] Lauren A Castro, Spencer J Fox, Xi Chen, Kai Liu, Steven E Bellan, Nedialko B Dimitrov, Alison P Galvani, and Lauren Ancel Meyers. As-

sessing real-time zika risk in the united states. *BMC infectious diseases*, 17(1):284, 2017.

[8] Centers for Disease Control and Prevention. Cdc's zika virus testing guidance. https://www.cdc.gov/zika/pdfs/testing-algorithm-symptomatic-nonpregnant.pdf. Accessed: 2018-09-04.

[9] Centers for Disease Control and Prevention. Zika mac-elisa. `https://www.fda.gov/downloads/medicaldevices/safety/emergencysituations/UCM488044.pdf`. Accessed: 2018-09-26.

[10] Centers for Disease Control and Prevention (CDC). Key facts about influenza (flu). `https://www.cdc.gov/flu/keyfacts.htm`. Accessed: 2017-01-09.

[11] Centers for Disease Control and Prevention (CDC). Trioplex real-time rt-pcr assay. `https://www.fda.gov/downloads/medicaldevices/safety/emergencysituations/ucm491592.pdf`. Accessed: 2018-09-26.

[12] Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Vector-Borne Diseases (DVBD). CDC Laboratory Guidance and Diagnostic Testing. `https://www.cdc.gov/dengue/clinicallab/laboratory.html`. Accessed: 2019-03-26.

[13] Xi Chen, Nedialko B Dimitrov, and Lauren Ancel Meyers. Uncertainty analysis of species distribution models. *PloS one*, 14(5):e0214190, 2019.

[14] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

[15] Nedialko B Dimitrov and Lauren Ancel Meyers. Mathematical approaches to infectious disease prediction and control. *JJ Hasenbein, ed. INFORMS TutORials in Operations Research*, 7:1–25, 2010.

[16] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.

[17] Miroslav Dudík, Steven J Phillips, and Robert E Schapire. Correcting sample selection bias in maximum entropy density estimation. In *Advances in neural information processing systems*, pages 323–330, 2005.

[18] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986.

[19] Jane Elith, Steven J Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J Yates. A statistical explanation of maxent for ecologists. *Diversity and distributions*, 17(1):43–57, 2011.

[20] Tomohiro Endo, Tomoaki Watanabe, and Akio Yamamoto. Confidence interval estimation by bootstrap method for uncertainty quantification using random sampling method. *Journal of Nuclear Science and Technology*, 52(7-8):993–999, 2015.

[21] Seasonal Influenza Flu. Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the united states. *Health Care*, 2007:2006–2007, 2008.

[22] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[23] Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F-Radar and Signal Processing*, volume 140, pages 107–113. IET, 1993.

[24] Norma Gorrochotegui-Escalante, Consuelo Gomez-Machorro, Saul Lozano-Fuentes, Lldefonso Fernandez-Salas, Maria De Lourdes Munoz, Jose A Farfan-Ale, Julian Garcia-Rejon, Barry J Beaty, William C Black, et al. Breeding structure of aedes aegypti populations in mexico varies by region. *The American journal of tropical medicine and hygiene*, 66(2):213–222, 2002.

[25] Antoine Guisan and Wilfried Thuiller. Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 8(9):993–1009, 2005.

[26] Rune Halvorsen, Sabrina Mazzoni, Anders Bryn, and Vegar Bakkestuen. Opportunities for improved distribution modelling practice via a strict

maximum likelihood interpretation of maxent. *Ecography*, 38(2):172–183, 2015.

[27] John M Hammersley and K William Morton. Poor man's monte carlo. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 23–38, 1954.

[28] Robert J Hijmans, Susan E Cameron, Juan L Parra, Peter G Jones, Andy Jarvis, et al. Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol*, 25(15):1965–1978, 2005.

[29] Yili Hong. On computing the distribution function for the poisson binomial distribution. *Computational Statistics & Data Analysis*, 59:41–51, 2013.

[30] Elizabeth A Hunsperger, Jorge Muñoz-Jordán, Manuela Beltran, Candimar Colón, Jessica Carrión, Jesus Vazquez, Luz Nereida Acosta, Juan F Medina-Izquierdo, Kalanthe Horiuchi, Brad J Biggerstaff, et al. Performance of dengue diagnostic tests in a single-specimen diagnostic algorithm. *The Journal of infectious diseases*, 214(6):836–844, 2016.

[31] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[32] Jagat Narain Kapur and Hiremaglur K Kesavan. *Entropy optimization principles and their applications*. Springer, 1992.

[33] Michael David Kavanaugh. Influence of stormwater drainage facilities on mosquito communities within the city of denton, texas. `http://digital.library.unt.edu/ark:/67531/metadc9765/m2/1/high_res_d/thesis.pdf`. Accessed: 2016-04-28.

[34] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[35] Minjung Kyung, Jeff Gill, Malay Ghosh, George Casella, et al. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

[36] Jorge M Lobo, Alberto Jiménez-Valverde, and Joaquín Hortal. The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1):103–114, 2010.

[37] Hedibert F Lopes and Ruey S Tsay. Particle filters and bayesian inference in financial econometrics. *Journal of Forecasting*, 30(1):168–209, 2011.

[38] Stéphanie Manel, Jean-Marie Dias, and Steve J Ormerod. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a himalayan river bird. *Ecological modelling*, 120(2):337–347, 1999.

[39] Bryan FJ Manly. *Randomization, bootstrap and Monte Carlo methods in biology*, volume 70. CRC Press, 2006.

[40] Lee P McPhatter, Farida Mahmood, and Mustapha Debboun. Survey of mosquito fauna in san antonio, texas. *Journal of the American Mosquito Control Association*, 28(3):240–247, 2012.

[41] Samuel A Merrill, Frank B Ramberg, and Henry H Hagedorn. Phylogeography and population strucure of aedes aegypti in arizona. *The American journal of tropical medicine and hygiene*, 72(3):304–310, 2005.

[42] Jennifer Miller. Species distribution modeling. *Geography Compass*, 4(6):490–509, 2010.

[43] Noelle-Angelique M Molinari, Ismael R Ortega-Sanchez, Mark L Messonnier, William W Thompson, Pascale M Wortley, Eric Weintraub, and Carolyn B Bridges. The annual impact of seasonal influenza in the us: measuring disease burden and costs. *Vaccine*, 25(27):5086–5096, 2007.

[44] Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937.

[45] Gabriela Paz-Bailey, Eli S Rosenberg, Kate Doyle, Jorge Munoz-Jordan, Gilberto A Santiago, Liore Klein, Janice Perez-Padilla, Freddy A Medina, Stephen H Waterman, Carlos Garcia Gubern, et al. Persistence of zika virus in body fluidspreliminary report. *New England Journal of Medicine*, 2017.

[46] Jennie L Pearce and Mark S Boyce. Modelling distribution and abundance with presence-only data. *Journal of applied ecology*, 43(3):405–412, 2006.

[47] Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3):231–259, 2006.

[48] Steven J Phillips, Miroslav Dudík, and Robert E Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning*, page 83. ACM, 2004.

[49] Michael K Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599, 1999.

[50] J Reiczigel, J Földi, and L Ózsvári. Exact confidence limits for prevalence of a disease with an imperfect diagnostic test. *Epidemiology & Infection*, 138(11):1674–1678, 2010.

[51] Ian W Renner and David I Warton. Equivalence of maxent and poisson point process models for species distribution modeling in ecology. *Biometrics*, 69(1):274–281, 2013.

[52] Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman filter: Particle filters for tracking applications*, volume 685. Artech

house Boston, 2004.

[53] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice hall Upper Saddle River, 2003.

[54] Samuel V Scarpino, Nedialko B Dimitrov, and Lauren Ancel Meyers. Optimizing provider recruitment for influenza surveillance networks. *PLoS Comput Biol*, 8(4):e1002472, 2012.

[55] Claude E Shannon. The mathematical theory of communication. 1963. *MD computing: computers in medical practice*, 14(4):306–317, 1996.

[56] Adrian FM Smith and Alan E Gelfand. Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

[57] Danny Soto. private communication (email), May 4, 2015. Environmental Services Department, El Paso.

[58] Michael Spivak. *Calculus on manifolds*, volume 1. WA Benjamin New York, 1965.

[59] Theodore E Sterne. Some remarks on confidence or fiducial limits. *Biometrika*, 41(1/2):275–278, 1954.

[60] MG Thompson, DK Shay, H Zhou, CB Bridges, PY Cheng, E Burns, JS Bresee, NJ Cox, et al. Estimates of deaths associated with seasonal

influenza-united states, 1976-2007. *Morbidity and Mortality Weekly Report*, 59(33):1057–1062, 2010.

[61] United States Census Bureau. American community survey (ACS). `https://www.census.gov/programs-surveys/acs/data.html`. Accessed: 2016-04-28.

[62] Rudolph Van Der Merwe, Arnaud Doucet, Nando De Freitas, and Eric Wan. The unscented particle filter. In *Nips*, volume 2000, pages 584–590. Denver, CO, USA, 2000.

[63] Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, and Antoine Flahault. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*, 158(10):996–1006, 2003.

[64] Christopher J Vitek, Joann A Gutierrez, and Frank J Dirrigl Jr. Dengue vectors, human activity, and dengue virus transmission potential in the lower rio grande valley, texas, united states. *Journal of medical entomology*, 51(5):1019–1028, 2014.

[65] A Yu Volkova. A refinement of the central limit theorem for sums of independent random indicators. *Theory of Probability & Its Applications*, 40(4):791–794, 1996.

[66] Gill Ward, Trevor Hastie, Simon Barry, Jane Elith, and John R Leathwick. Presence-only data and the em algorithm. *Biometrics*, 65(2):554–

563, 2009.

[67] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[68] Wikipedia. Poisson binomial distribution. `https://en.wikipedia.org/wiki/Poisson_binomial_distribution`. Accessed: 2018-09-05.

[69] World Health Organization. Dengue and severe dengue. `https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue`. Accessed: 2019-06-3.